COMPUTERWORLD

CIO BUSINESS WORLD

Bigdata 2013

Neztraťte se v záplavě dat

HUAWEI

sas

T · · Systems ·

TERADATA

THE BEST DECISION POSSIBLE™

Obsah

4 Postupy, díky kterým své projekty big dat nasadíte úspěšně

17 Když Hadoop už nestačí...

7 Big data v malém Česku

8 Ideální řešení pro big data?

11 Big data v cloudu jsou se SAP realita

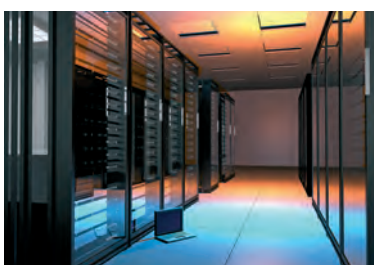
12 Budoucnost BI je v rychlém zpracování dat

13 Vidět budoucnost je lepší než jen tušit

20 Máte včas podklady pro kvalifikovaná rozhodnutí?

22 Cesta k efektivnímu zhodnocení podnikových informací

24 NoSQL: Ne pro všechno se hodí



14 U big dat nestačí jen nakoupit podstatně větší úložiště

27 Big data usnadňují bezpečnostní korelace

COMPUTERWORLD CIO BUSINESS WORLD

UCELENÝ INFORMAČNÍ ZDROJ PRO IT PROFESIONÁLY

IDG INTERNATIONAL DATA GROUP FIPP Člen asociace FIPP

Vydává: IDG Czech Republic, a. s., Seydlerova 2451, 158 00 Praha 5
Tel. ústředna s aut. provolbou: 257 088 + linka; fax: 235 520 812
Recepce: 257 088 111
Výkonný ředitel: RNDr. Jana Pelikánová
Šéfredaktor: Radan Dolejš
Tajemnice redakce: Růžena Holíková, tel.: 257 088 143
Vedoucí inzertního odd.: Jitka Vyhlídková, tel.: 257 088 181
Vedoucí projektu: Pavel Louda, tel.: 257 088 138
Jazyková úprava: Dana Štropová
Obálka: Petr Kubát
Adresa redakce: CW, Seydlerova 2451, 158 00 Praha 5

Internet: jmeno_prijmeni@idg.cz
Zlom a pre-press: TypoText, s. r. o., Praha
Tisk: Libertas, a. s.
Předplatné a reklamace: IDG Czech Republic, a. s., Seydlerova 2451, 158 00 Praha 5, tel.: 257 088 163, fax 235 520 812; e-mail: predplatne@idg.cz
Doručuje Česká pošta, s. p., v systému D + 1
Předplatné pro Slovensko: Magnet-Press Slovakia, s. r. o., P.O.BOX 169, 830 00 Bratislava, tel.: +421 267 201 910, 20, 30, e-mail: predplatne@press.sk

Copyright: © 2013 IDG Czech Republic, a. s.

Nezůstaňte stranou

Když se platforma Hadoop v posledních měsících intenzivně skloňuje zejména v souvislosti se vzestupem popularity řešení big dat, experti očekávají, že se v nejbližších letech rozšíří v souvislosti s klasickými firemními analýzami i do standardního firemního IT.

PAVEL LOUDA,
VEDOUCÍ PROJEKTU

Přestože dosavadní implementace Hadoopu přinesly zvýšené nároky na dovednosti příslušných správců, firmy si podle Gartneru slibují díky širšímu využití platformy kvalitnější výsledky rozborů svých dat, zejména málo strukturovaných, textů či údajů ukazujících vzorce nějakého chování.

Dá se tedy očekávat, že s Hadoopem budou firmy v následujících letech hojně experimentovat – zejména v oblasti appliancí a klasických databází.

Rovněž dodavatelé analytických systémů se ve vyšší míře zaměří na vestavěné funkce svých analytických systémů, které budou obsahovat prvky Hadoopu. Podle Gartneru by tak mohly fungovat už během pouhých dvou let až dvě třetiny klasických aplikací s rozšířenou analytikou.

Tato změna bude mít samozřejmě vliv i na chod firemních IT oddělení – ta budou muset zajistit, aby podniková infrastruktura byla s Hadoopem kompatibilní, a analytická oddělení zase budou nucena vytvářet analytické funkce, které jsou pro byznys velmi relevantní – právě za pomoci projektů s podporou Hadoopu.

Vývoj analytických nástrojů však podle Gartneru půjde ještě dál – už za pár let většina příslušných dodavatelů přidá do svých řešení také podporu přirozeného jazyka či prvky pro hlasové ovládání – cílem je samozřejmě maximálně zjednodušit práci s těmito systémy, aby i běžný uživatel včetně mobilních dokázal s analytickými řešeními efektivně pracovat.

Jak je vidět, v oblasti správy velkých objemů dat i pokročilých analýz nás v následujících měsících a letech čeká obrovský kvalitativní skok. Je jen na každém, zda se mu dokáže včas přizpůsobit.

A k tomu, jak věříme, může přispět i tato publikace, která vás s mnohými taji nových technologií práce s objemnými daty seznamuje. Příjemné a poučné počtení. ■

EMC² hp VERTICA HUAWEI SAP
sas T-Systems TERADATA THE BEST DECISION POSSIBLE™

EMC²
hp
VERTICA
HUAWEI
SAP®

Postupy, díky kterým své projekty big dat nasadíte úspěšně

Nenaplněná očekávání, zvýšené náklady, zbytečná právní rizika – vkročit do světa big dat naslepo se nevyplácí.

BOB VIOLINO

Technologie big dat se objevila teprve nedávno jako velká iniciativa v podnikové sféře. Moderní, experimentální a svým způsobem účelové pozlátko směsi datových proudů s odhalováním významných skutečností však neodpovídá obrovským kulturním a technologickým nárokům, na které zatím mnoho podniků není připravených. Bez strategického plánu, který zahrnuje promyšlené cíle, velmi dobré řízení dat, pečlivé procesy pro zajištění přesnosti údajů, správného myšlení a lidí, mohou iniciativy mající v hledáčku big data snadno skončit jako fiasko namísto toho, že budou mít pro organizaci skutečně cenný přínos.

Přinášíme pět strategických tipů, jak se ochránit před neúspěchem při nasazení big dat. V mnoha případech se doporučení vztahují na jakýkoliv projekt správy dat nezávisle na objemu datové množiny.

Příchod obrovských datových skladů však s sebou přinesl určitou řadu nástrah. Zde je návod, jak zvýšit šance, aby úsilí vaší organizace míchat velké objemy dat z různých zdrojů bylo nakonec úspěšné.

TIP Č. 1 PRO ÚSPĚCH BIG DAT

Udělejte z big dat hlavní firemní strategii

Phil Steitz, technologický ředitel společnosti Rearden Commerce, stručně shrnuje nejdůležitější faktor úspěchu big dat: Musíte integrovat analytická a daty řízená rozhodování přímo do jádra své firemní strategie.

„Pokud ale budou big data ve vaší organizaci jen populárním slovem, koleduje si takové pokusné řešení o velké problémy,“ varuje Steitz.

Pro firmu Reardon Commerce, jejíž e-commerce platforma využívá big data a další zdroje pro optimalizaci výměny zboží, služeb a informací mezi kupujícími a prodávajícími, má tento koncept „absolutní důležitost“. Klíčem je prý předložit vhodnou obchodní příležitost ve správný okamžik vhodnému ekonomickému zprostředkovateli.

„Je to příklad tohoto druhu myšlení – vytvoření a centrální realizace obchodní strategie samotným vedením firmy,“ uvádí Steitz.

Součástí tohoto přístupu je vytvoření malého a vysoce schopného týmu datových vědců, sémantických analytiků či inženýrů big dat a zahájení nepřetržitého obousměrného dialogu mezi tímto týmem a prozíravými obchodními manažery, kteří řídí firmu a rozhodují o zakázkách, popisuje Steitz.

„Největším problémem při opravdovém získávání hodnoty IT technologií nasazených pro analytiku a sémantické rozbory je, aby technologové, kteří skutečně dokážou vytěžit nějaké hodnoty, dostatečně spolupracovali s firemními manažery, kteří ale výsledky musí umět využít a vybrat to, co je opravdu cenné,“ vysvětluje Steitz.

Dalším klíčovým faktorem úspěchu při integraci big dat do celkové firemní strategie je efektivní správa takzvaných datových partnerství.

„Skutečná optimalizace zákaznické zkušenosti a ekonomické hodnoty v dnešním světě nevyhnutelně vyžadují sdílení dat v rámci celého podniku,“ prohlašuje Steitz a pokračuje: *„Naivní přístupy k této problematice způsobem ‚pošlete nám přes noc celý soubor transakcí‘ končí naprostým selháním jak z provozního hlediska, tak z pohledu ochrany soukromí a zabezpečení.“*

TIP Č. 2 PRO ÚSPĚCH BIG DAT

Správa dat je nepostradatelná

Projekty spojené s big daty s sebou nesou poměrně závažné starosti související se zabezpečením, soukromím a dodržováním předpisů či jiných legislativních norem. Nejcitlivější záležitostí je to ve sféře zdravotnických služeb.

Beth Israel Deaconess Medical Center, významný poskytovatel zdravotní péče, je jednou z organizací, které se začínají o big data poměrně značně zajímat. Pracují tam totiž s elektronickými zdravotními záznamy, novými modely úhrady zdravotní péče a obrovským množstvím klinických dat a požadavků nasbíraných za léta provozu. Řízení dat v tomto případě bude hrát klíčovou roli.

„Na IT oddělení poskytovatelů zdravotní péče se bude vyvíjet velký tlak, aby rychle zpřístupňovala svá data oprávněným osobám,“ tvrdí Bill Gillis, tamější ředitel IT. Implementace solidního řízení organizacím umožní zajistit, aby tyto údaje byly přesné a podávaly klinické informace potřebné pro poskytování kvalitní a stále se zlepšující péče.

„Je velmi důležité, aby nezvítězila ‚tyranie naléhavosti‘,“ zdůrazňuje Gillis a dodává: *„Implementace způsobu řízení ‚předem‘ může pomoci předejít nástrahám a udržet všechny záležitosti na správné cestě.“*

Samozřejmě že zabezpečení a soukromí jsou toho významnými součástmi. *„Vzhledem k nejistotám, které obklopují big data, je pro důležité značky laťka soukromí a zabezpečení tak vysoko, že ochrana poskytovaná pro tyto nové technologie je větší než u většiny ostatních tradičních systémů pro rozhodování,“* říká Charles Stryker, předseda a výkonný ředitel firmy Venture Development Center.

Tato společnost poskytuje konzultační služby a další rady v oblasti big dat pro organizace jako AOL, Cisco, First Data nebo Yahoo. *„Žádná významná značka nechce testovat hranice, kde už končí soukromí a zabezpečení,“* vysvětluje Stryker.

Pokud firmy chtějí použít nějaká data v jakémkoliv řešení big dat, musí hned na začátku projektu zvážit jejich původ (metadata, která popisují zdroj dat) a udělat vhodná rozhodnutí týkající se jejich zařazení (citlivost, důvěrnost dat), myslí si Louis Chabot, hlavní technický poradce a lídr pro big data u konzultační společnosti DRC, která působí v oblasti technologií a managementu a pomáhá organizacím implementovat projekty big dat.

„Údržba metadat zahrnujících informace o původu dat a rozhodování o jejich zařazení nejsou záležitost, kterou by šlo řešit až zpětně,“ varuje Chabot a dodává: *„Jsou to nedílné součásti iniciativy, která se musí navrhnout a zahrnout už od samého počátku projektu.“*

Pokud je to vhodné, tvrdí Chabot, měly by se k ochraně před náhodným nebo zlovolným pozměněním údajů použít specializované metody, jako jsou například digitální podpisy.

Organizace také musí respektovat zákony a předpisy na ochranu osobních údajů. Aby dokázaly splnit požadavky, jež kladou předpisy a normy pro ochranu osobních údajů a zabezpečení dat, *„mohou firmy použít různé metody – jsou jimi například anonymizace údajů, oddělování částí dat či omezení distribuce a využití některých typů informací,“* uvádí Chabot.

TIP Č. 3 PRO ÚSPĚCH BIG DAT

Nepodceňujte správnost dat

Nedávný výzkum agentury Aberdeen Group zdůrazňuje ještě další aspekt potřebný pro úspěch big dat: správnost dat.

Společnosti, jež jsou podle metrik Aberdeenu nejlepší ve své kategorii, uvádějí, že jejich organizačním cílem byla 94% správnost dat – pro splnění tohoto cíle tak musejí zajistit zlepšení o jedno procento.

Oborově průměrné firmy ohlásily jako cíl správnosti dat hranici 91 procent – aby toho dosáhly, potřebují zlepšení svých metodik pro správu dat o 18 %. Zaostalejší společnosti hlásily cíl přesnosti dat 80 %, a tudíž nutnost 40% zlepšení svého současného výkonu.

Pro úspěch projektů big dat jsou zcela zásadní čištění dat a práce s nejdůležitějšími údaji pro provoz firmy.

„Na rozdíl od některých jiných požadavků právě tento nezmizí,“ uvádí Joe Caserta, zakladatel a výkonný ředitel společnosti Caserta Concepts, která poskytuje konzultace pro správu klasických firemních dat a big dat.

„Aby se model big dat hodil jako nová platforma firemních analýz, musí být tato technologie schopná přizpůsobit se zákazníkům, produktům, zaměstnancům, umístění atd, a to bez ohledu na konkrétní zdroj informací,“ tvrdí Caserta.

Kromě toho budou mít na analýzy big dat stejný vliv známé problémy s kvalitou údajů, které dlouho ohrožovaly věrohodnost analýz, pokud se vhodně nevyřeší, připomíná Caserta.

U obvyklého současného projektu big dat se správa údajů často podceňuje a má u vývojového týmu nízkou prioritu, takže nakonec může zůstat nedořešená, poznamenává Chabot z DRC.

Efektivní správa dat zahrnuje použití vyspělých metod – procesů a automatizace – pro zajištění správy modelu, metadat, referenčních údajů, kmenových dat (MDM, Master Data Management), slovní zásoby, kvality údajů a inventáře informací, dodává Chabot.

TIP Č. 4 PRO ÚSPĚCH BIG DAT

Shromažďujte osvědčené postupy k dosažení nejlepších výsledků

Při správě big dat a při odpovídajících analýzách lidé zjišťují, co funguje a co ne. Když jsou zaměstnáni ve stejné firmě, proč by se neměli o tyto poznatky dělit s ostatními pracovníky?

Jedním způsobem, jak to udělat, je vytvořit takzvanou big data COE (Centre of Excellence) – sdílenou entitu, která bude poskytovat informace z řízení, osvědčené postupy a v některých případech také podporu či školení.

„Střediska COE mají vlastní vyhrazený rozpočet a jsou navržená tak, aby analyzovala problémy, definovala různé iniciativy, budoucí (žádoucí) stav a standardy, dále školila uživatele, dělala plány a zajišťovala modernizaci a aktualizaci řešení,“ tvrdí Eliot Arnold, spoluzakladatel poradenské firmy Massive Data Insight, která se specializuje na big data a analytické programy.

Založení centra COE podle něj ale vyžaduje audit dostupných zdrojů a podporu vyššího managementu organizace.

Přestože je centrum COE pro projekty big dat dobrý nápad, bude jeho celkovou efektivitu určovat především to, jak dobře bude v praxi fungovat, upozorňuje Chabot z DRC.

Existuje totiž celá řada základních problémů týkajících se pokrytí celého životního cyklu dat pomocí COE, připomíná Chabot – včetně autorství a identifikace osvědčených postupů, jejich prověřování nezaujatým způsobem (třeba třetí stranou), vhodné dokumentace použitelnosti, dohledu nad nasazením nebo aktualizace v průběhu času.

Firma DRC definovala úroveň vyzrálosti big dat podobně jako CMMI (Capability Maturity Model Integration), což je firmami využívaný model zralosti pro zlepšování procesů.

Typy úrovně zralosti big dat mapují relevantní osvědčené postupy.

Rozdělují se do čtyř skupin: plánování/správa, uskutečnění projektů, architektura a nasazení/provoz/realizace. Organizace je mohou zavádět postupnými kroky. *„To zabrání nebezpečí spočívajícímu v pokusech o velmi rychlé dosažení přílišné propracovanosti,“* dodává Chabot.

TIP Č. 5 PRO ÚSPĚCH BIG DAT

Klíčové jsou odbornost a vzájemná spolupráce

Big data představují celofiremní iniciativu – není to tedy jen technologický projekt, takže je nevyhnutelné, aby vedení firmy i IT oddělení kooperovala při plánování, realizaci a údržbě projektu.

„Jedním z největších nebezpečí pro program big dat je nedostatečná spolupráce mezi vedoucími pracovníky IT a firemním managementem, který řídí komplexní strategie a iniciativy,“ upozorňuje Arnold.

„V méně vyspělých organizacích neexistuje zdokumentovaná strategie, v rámci provozu se využívá směsice nástrojů a osoby s rozhodovací pravomocí upřednostňují při určování strategického směru intuici. Tyto typy firem si většinou hodnotu dat nijak zvlášť neuvědomují,“ tvrdí Arnold.

Firemní vedoucí představitelé a odborníci na oblast IT mohou zajistit úspěšnost svých projektů big dat pouze pečlivým stanovením cílů, potřeb a požadavků, výpočtem návratnosti investic, přiřazením analytických schopností k potřebám firmy a zavedením mechanismů pro průběžnou zpětnou vazbu, radí Chabot z DRC.

„Projekt big dat by se měl rozdělit do více fází, které budou postupně zvyšovat hodnotu nového řešení pro samotnou organizaci,“ doporučuje Chabot.

Zajistit harmonickou spolupráci vedení firmy a IT stejně jako kooperaci jednotlivých oddělení na datových iniciativách však nebývá snadné.

„Podle mých zkušeností se to stává ve velkých společnostech skutečnou korporátní výzvou,“ uvádí Stryker z Venture Development Center.

„Má práce ředitele pro oblast podnikových dat spadat, co se týče odpovědností, do oddělení IT, marketingové sekce, divize řízení rizik, anebo mají mít všechna tato oddělení své vlastní iniciativy pro big data a vzájemně je jen koordinovat,“ ptá se Stryker.

Společnosti také musí využít pro nasazení technologií big dat nezbytný odborný personál – jako třeba v případě Hadoopu. Jen to jim umožní mít levnou a výpočetně efektivní správu velmi velkých datových množin a analytických úloh.

„Posun paradigmatu směrem k technologii big dat přináší do prostředí podniků naprosto novou roli – datové vědce,“ upozorňuje Caserta.

„Tato úloha vyžaduje hluboké znalosti vyšší matematiky, systémového a datového inženýrství a v neposlední řadě také obchodní odborné schopnosti.“

V praxi je běžné využívat celý vědecký datový tým tvořený statistiky, technology a experty na předmět firemního podnikání, jež kolektivně řeší problémy a poskytují řešení, popisuje situaci Caserta a dodává, že mnoho lidí, kteří dnes pracují v oblasti analýz dat, se musí připravit na kulturní šok.

„Před zahájením projektu big dat je potřeba udělat strategický test připravenosti ohledně posouzení přijetí nového paradigmatu,“ radí Caserta. Obchodní analytici budou muset projít přeškolením nebo změnit své zaměření.

Cíl přechodu na platformu big dat může také zahrnovat přechod od reaktivní analýzy (například jak dobře fungovala reklamní či jiná kampaň) na prediktivní (co by měla příští kampaň obsahovat).

Tak nyní můžeme, jak popisuje Caserta, aktivně ovlivňovat zatím nekupující klienty, aby se začali chovat podle vzorců loajálních zákazníků, nebo znovu stimulovat aktivní klienty, pokud jejich vzorce chování začínají napovídat, že je jako zákazníky ztrácíme.

A jaká jsou rizika spočívající v tom, že se nevytvoří silné a kompaktní strategie big dat? Nákladný pokus, který nepřinese očekávaný užitek.

„Projekty big dat jsou obvykle vícerozměrné a složité iniciativy,“ říká Chabot a dodává: *„Vyžadují od odpovědných osob značné předchozí plánování.“*

Před zahájením projektu big dat by mělo vedení firmy zajistit soulad mezi strategickými, funkčními, datovými, analytickými a technologickými plány. Ty se přitom musí zohlednit ve firemní, systémové, softwarové, datové a technologické architektuře.

„Nesoulad mezi některými z těchto plánů může způsobit vykolejení celého projektu,“ varuje Chabot.

„Rizika neexistující silné a kompaktní strategie big dat s odpovídajícími plány a architekturami budou pravděpodobně zahrnovat jevy, jako nepřiměřené náklady, nesourodá očekávání, nedostatek přidané hodnoty, a nakonec mohou vést až k celkovému selhání projektu,“ uzavírá Chabot. ■

Big data v malém Česku

Technologie pro zpracování velkých objemů dat najde své uplatnění i v malé zemi, jakou je Česko.

LUDĚK ŠAFÁŘ

Předpokládám, že zhruba tušíte, co je problematika big dat. Nejen my z EMC, ale i naši konkurenti vás v této oblasti pravidelně zásobují mnoha víceméně podstatnými informacemi, a to nejen na stránkách této publikace. Velké množství těchto informací má původ tam, kde většina novinek v oblasti IT – v zámoří. Ale jsou vůbec hlavní směry využití technologií velkých dat aplikovatelné na rozměry, se kterými se potkáváme u nás?

Co funguje a co ne

Obecná odpověď zní, že ano. Ona totiž základní myšlenka výše zmíněného konceptu (vytěžování důležitých informací z velkého množství „obyčejných" dat) není nijak nová.

I proto tak často odpovídáme našim zákazníkům, kteří jsou v této oblasti už dlouho aktivní, na otázku, „co se najednou tak změnilo"? Změnilo, čtěte dál.

Ale nejdřív trochu střízlivění. Některé aplikace big dat totiž v Česku nefungují a pravděpodobně ani jen tak nebudou. Typickou ukázkou je velmi často zmiňované vytěžování dat ze sociálních sítí, zejména Twitteru.

I když počet uživatelů Twitteru v ČR od roku 2008 víceméně lineárně roste, stále je aktivních uživatelů cca 8 % populace (150 tisíc), což velmi pravděpodobně ani z demografického hlediska nemůže představovat reprezentativní vzorek.

To ostatně platí nejen u nás – rozdělení uživatelů Twitteru je velmi nerovnoměrné, geograficky, i co se týká vzdělání, zájmů apod. A jak se ukazuje, i zpravodajská relevance Twitteru se mnohdy přeceňuje (většina tweetů týkajících se hurikánů přichází z oblastí, kde hurikán není). Twitter zmiňuji zcela záměrně, protože jde o typický příklad aplikace big dat na „lidi" – uživatele, zákazníky.

Samozřejmě existují velmi typické okruhy, kde naopak vytěžování dat funguje – chování spotřebitelů, klientů bank a podobně – pro tyto organizace jsou také technologie datových skladů jako předchůdce big dat denním chlebem.

Co je jinak?

Technologie ale pokročila, a to zásadně. Zrychlení původně časově náročných analýz z hodin na minuty a z minut na vteřiny umožňuje jejich využití ve zcela nových scénářích. Během čtení webové stránky. Během rozhovoru s operátorem. Během placení u pokladny. A nové scénáře znamenají další obchodní modely i tržní příležitosti.

Koncepčně nový způsob využití analytických principů je v oblasti týkající se dat z jiných okruhů než chování uživatelů a zákazníků – například ve výrobě, kde dochází ke shromažďování velkého množství dat a nabízí se jejich korelace k atributům spojeným s finálním výrobkem – spolehlivosti, kvality apod.

Současně může docházet ke zvětšování relevantní datové báze, která do analýzy vstupuje. Zvětšování například v takovém rozsahu, že neumožňuje dopředu odhadovat, jaké můžou být výsledky těchto analýz. Začíná se ukazovat, že data se vyplácí shromažďovat a zkoumat. Obecně se předpokládá, že jde o zatím nevyužitý zdroj reálných obchodních příležitostí, jen ne vždycky patřičně vytěžený.

V neposlední řadě je třeba mít na paměti, že s nárůstem dostupného výkonu jde ruku v ruce i snížení nákladů, které pro big data potřebujete. Základní koncepty můžete vyzkoušet bez investice do drahého hardwaru – stáhněte si Greenplum software a experimentujte. A pokud nejste v big datech žádní nováčci – nechte si nacenit Greenplum jako alternativu. Nebo si vybudujte paralelní analýzu, skoro zadarmo!

Pivotal – váš pivot

EMC, společně se společnostmi VMware a GE, v současné době spouští novou iniciativu Pivotal (*www.gopivotal.com*), která v sobě shrnuje všechny podstatné komponenty, jež s technologií big dat souvisejí. Jak produktové portfolio Greenplum (databáze, enterprise Hadoop), tak nové aktivity, které úzce souvisejí s analytickými nástroji a softwarovými moduly umožňujícími prezentaci dat a jejich zkoumání.

Autor je Senior Systems Engineer Manager, EMC Czech Republic s.r.o., *www.emc.com*

EMC FORUM 2013
LEAD YOUR TRANSFORMATION
9. září 2013
Clarion Congress hotel, Praha
EMC²

Ideální řešení pro big data? Huawei OceanStor

Nabídka úložných řešení společnosti Huawei zahrnuje diskové systémy, SAN i NAS platformy, virtuální páskové knihovny i speciální řešení pro ukládání dat v cloudu. Vysoce výkonná a spolehlivá disková úložiště Huawei jsou efektivním řešením centralizovaných datových skladů s vysokou dostupností a snadným sdílením dat.

JAROSLAV SITA

Zákazníci velmi dobře hodnotí produktovou řadu Huawei OceanStor T. Produkty této série mají vynikající poměr ceny a výkonu. Jde o modely S2600T/S5500T/S5600T/S5800T. Na jednom zařízení nabízejí jak blokový (SAN), tak souborový (NAS) přístup k datům, jde tedy o unified storage. Jsou široce škálovatelné (maximum je 1 440 disků v jednom diskovém poli), podporují všechny potřebné protokoly (iSCSI, FC, FCoE, NFS, CIFS, HTTP a FTP), úrovně RAID (0, 1, 3, 5, 6, 10, 50) a mají pokročilé softwarové funkce, jako jsou snapshot, SSD cache, thin provisioning, asynchronní i synchronní replikace a další. Konfigurují a spravují se pomocí velmi intuitivního grafického centrálního managementu, jsou tedy nenáročné na školení administrátorů, což výrazně šetří náklady a snižuje celkové TCO.

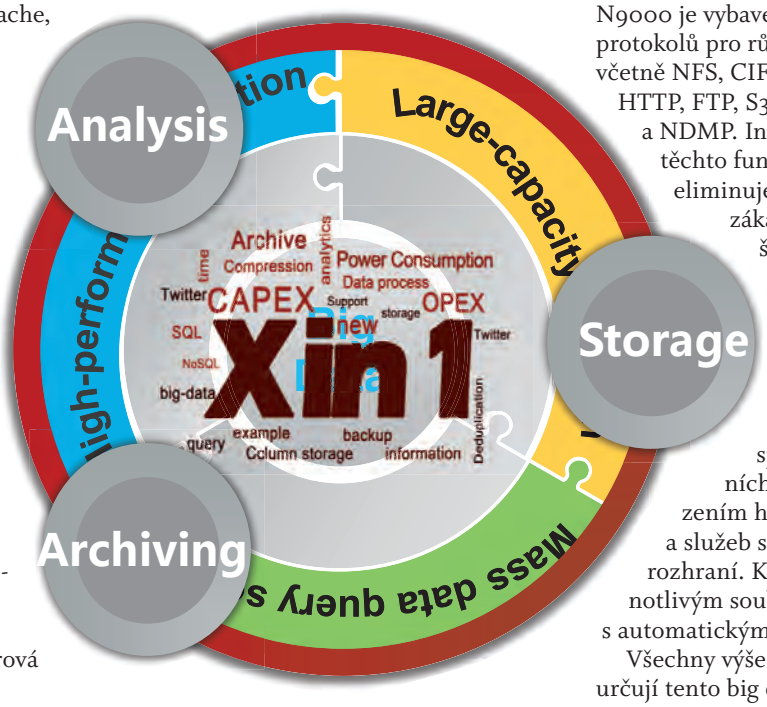
Dalšími velmi zajímavými produkty jsou OceanStor Dorado 2100 G2 a 5100. Vyznačují se fantastickou výkonností. Jsou totiž osazeny pouze SSD disky, mají pokročilý cache management a I/O algoritmus, což zajišťuje vysokou rychlost a nízkou latenci. Například u Dorado5100 bylo v SPC-1 naměřeno 600,052. 49. Ve druhé polovině letošního roku nabídneme u existujících produktových řad nové modely a různá hardwarová a softwarová vylepšení.

Podle různých průzkumů se odhaduje, že nestrukturovaná data představují 85 až 90 % informací ukládaných ve firmách a organizacích. Analytici předpovídají jejich masivní nárůst v příštích letech, který překoná již i tak velká očekávání.

Je tedy potřeba zajistit jejich efektivní ukládání i zálohování, pro což není sdílený prostor na serveru vůbec vhodný. Na rozdíl od toho systémy typu NAS poskytují vysoký výkon, snadnou rozšiřitelnost úložné kapacity, jednotnou administraci a též deduplikaci, která dokáže významně ušetřit diskový prostor. Jsou ideální volbou pro efektivní správu a sdílení velkých objemů dat. Všechny tyto požadavky a předpoklady splňuje nový produkt big data storage systém OceanStor N9000.

Big data storage systém OceanStor N9000

Huawei OceanStor N9000 storage systém si klade za cíl zjednodušit a splnit požadavky kladené při výběru úložných systémů.



N9000 je prvním ze storage systémů, který nabízí unifikovanou a konvergovanou infrastrukturu pro škálovatelné NAS systémy, databáze a backup řešení určených pro big data a podnikový segment.

Škálovatelnost

Poskytuje distribuovanou, plně škálovatelnou a vysoce spolehlivou architekturu bez jediného slabého místa. N9000 umožňuje připojení 3 až 288 nodů a až 100 PB v jednom file systému.

Vynikající výkon

Vyrovnávací paměť pro čtení a zápis dat je mnohem rychlejší než u tradičních pevných disků. N9000 používá globální architekturu vícevrstvých služeb s vyrovnávací pamětí 2 TB, což výrazně zlepšuje poměr přístupů a celkový výkon systému. N9000 dále ukládá metadata pomocí SSD disků pro zvýšení výkonu a přístupu k řízení kritických dat a zvyšuje schopnost zpracování malých souborů. N9000 využívá 10GE/infiniBand pro vysokorychlostní připojení k síti.

Dále je tento storage systém špičkou ve své oblasti díky rychlosti tři miliony operací za sekundu (OPS), využití 170 GB z celkové šířky pásma a velmi nízké latenci.

Inteligentní konvergence

N9000 je zařízení, které poskytuje unifikovanou a konvergovanou infrastrukturu pro škálovatelné NAS systémy, databáze a backup řešení. Tím podporuje jak strukturovaná, tak nestrukturovaná data. N9000 je vybaven řadou rozhraní a sadou protokolů pro různé přístupy k datům včetně NFS, CIFS, SMB v1, v2 SMB, HTTP, FTP, S3, JDBC, ODBC SQL92 a NDMP. Inteligentním zahrnutím těchto funkcí a protokolů N9000 eliminuje nedostatky a představuje zákazníkům komplexní řešení, které je ideální pro oblast big dat.

Zjednodušená správa

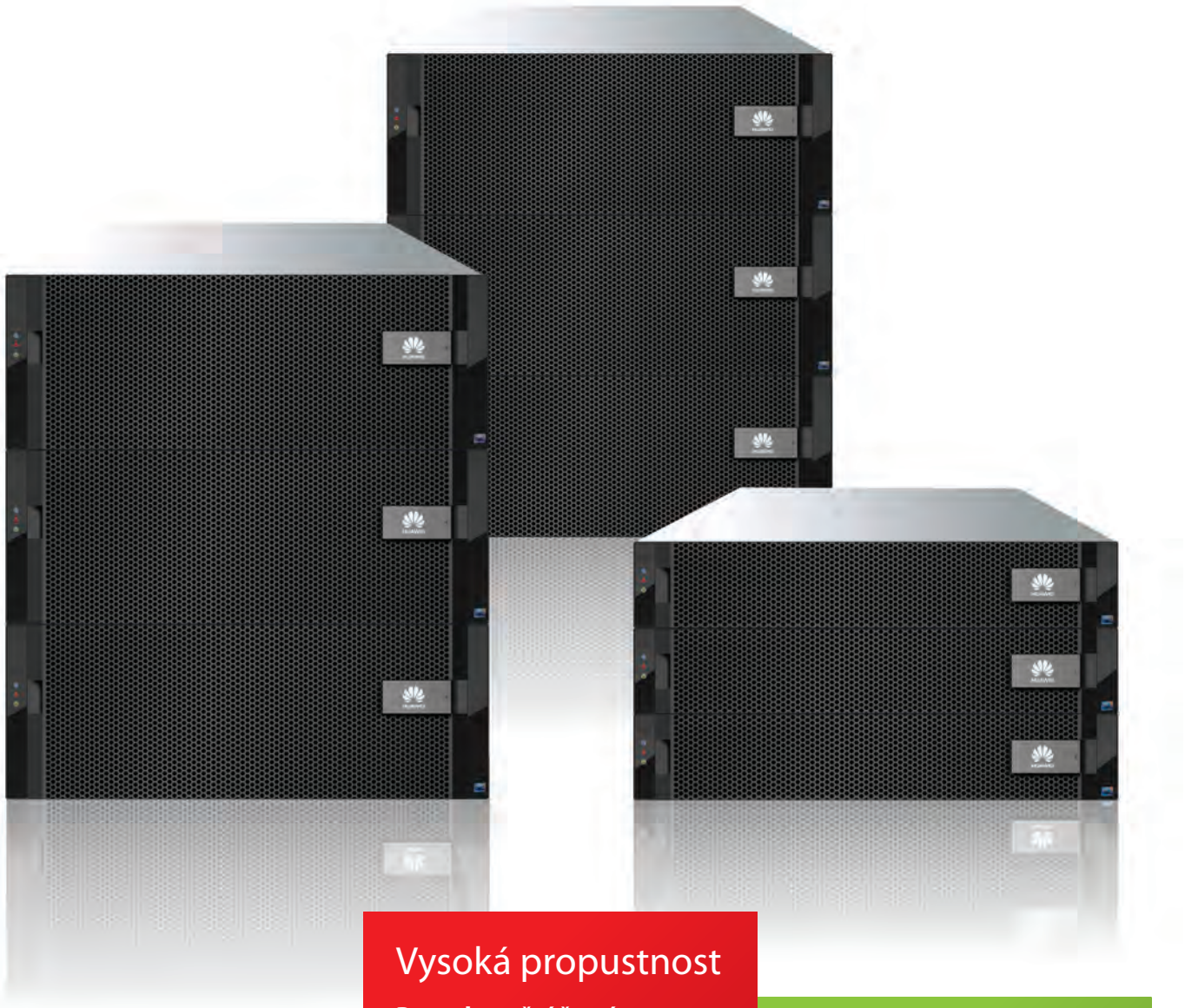
Dynamický systém řízení šetří čas tím, že sjednotí řešení komplexních úloh, a to centrálním řízením hardwaru, softwaru, sítě a služeb skrze jedno unifikované rozhraní. Kromě toho přiřazuje jednotlivým souborům jmenný prostor s automatickým tenkým provisioningem.

Všechny výše uvedené vlastnosti předurčují tento big data storage systém Huawei OceanStor N9000 k neuvěřitelně široké možnosti využití při různých scénářích.

Autor je Product manager ve společnosti DNS

HUAWEI
Enterprise Partner
Distributor

HUAWEI OceanStor N9000
Storage System



Velká kapacita
- Výška 4 U
- 36 diskových slotů
- 2-kanálové 4-jádrové CPU
- 16 GB paměti

Vysoká propustnost
Pro sekvenční čtení a zápis velkých souborů
- Výška 4 U
- 6 diskových slotů
- 2-kanálové 6-jádrové CPU
- 48 GB paměti

Vysoký výkon
Pro časté čtení a zápis malých souborů
- Výška 2 U
- 25 diskových slotů
- 2-kanálové 6-jádrové CPU
- 48 GB paměti
- SSD pro ukládání metadat

dns

www.dns.cz

ICT ve zdravotnictví

ICT VE ZDRAVOTNICTVÍ 2013

18. ZÁŘÍ 2013, MALOSTRANSKÝ PALÁC, PRAHA

7. ročník odborné konference a výstavy o elektronickém zdravotnictví, informačních a komunikačních technologiích

STÁT, LEGISLATIVA, VZDĚLÁVÁNÍ

EFEKTIVNÍ E-HEALTH

ELEKTRONICKÉ ZDRAVOTNICTVÍ, STANDARDY

ZDRAVOTNÍ DOKUMENTACE, IDENTIFIKACE, PLATEBNÍ SYSTÉMY

ZDRAVOTNÍ INFORMAČNÍ A KOMUNIKAČNÍ SÍTĚ, KOMUNIKACE A TECHNOLOGIE

ELEKTRONICKÉ REGISTRY, DATABÁZE

ZDRAVOTNICTVÍ A INFORMAČNÍ SPOLEČNOST

TELECARE – ASISTOVANÁ ZDRAVOTNÍ A SOCIÁLNÍ PÉČE

IDG
INTERNATIONAL DATA GROUP

www.idg.cz, www.eventworld.cz

Big data v cloudu jsou se SAP realita

Cloud computing je v současnosti vedle big dat asi druhým největším byznysovým a IT trendem. Bylo tedy jen otázkou času, kdy se na trhu objeví řešení spojující tyto dvě oblasti do jediné. A s ním přišla společnost SAP. V reakci na poptávku zákazníků oznámila dostupnost služby SAP HANA Enterprise Cloud.

Nyní tak lze využívat SAP HANA, in-memory platformu pro práci s big daty, také jako službu. Tato nabídka umožňuje organizacím nový způsob nasazení SAP HANA, kdy mohou prakticky ihned získat přínos z jejích inovativních vlastností.

Kritické podnikové systémy jako SAP ERP, SAP CRM, SAP NetWeaver Business Warehouse a řadu další aplikací nyní společnost SAP poskytuje i v podobě řízené cloudové služby s obrovskou, až petabajtovou škálovatelností, navíc s nízkými celkovými náklady na vlastnictví (TCO), flexibilitou a spolehlivostí, kterou je společnost SAP jako přední dodavatel podnikových aplikací schopna poskytnout. Cena přitom závisí především na rozsahu zpracovávaných dat a používaných aplikací.

Vlajkovou lodí je pro SAP sada aplikací SAP Business Suite, jež je nově postavena na platformě SAP HANA. Tato podniková řešení fungující nad daty v reálném čase lze nyní také využívat v rámci SAP HANA Enterprise Cloud.

Společnost SAP je jediným poskytovatelem integrované sady podnikových aplikací, které mohou přistupovat k transakčním datům a analyzovat je v reálném čase a na jednotné platformě pracující v režimu in-memory.

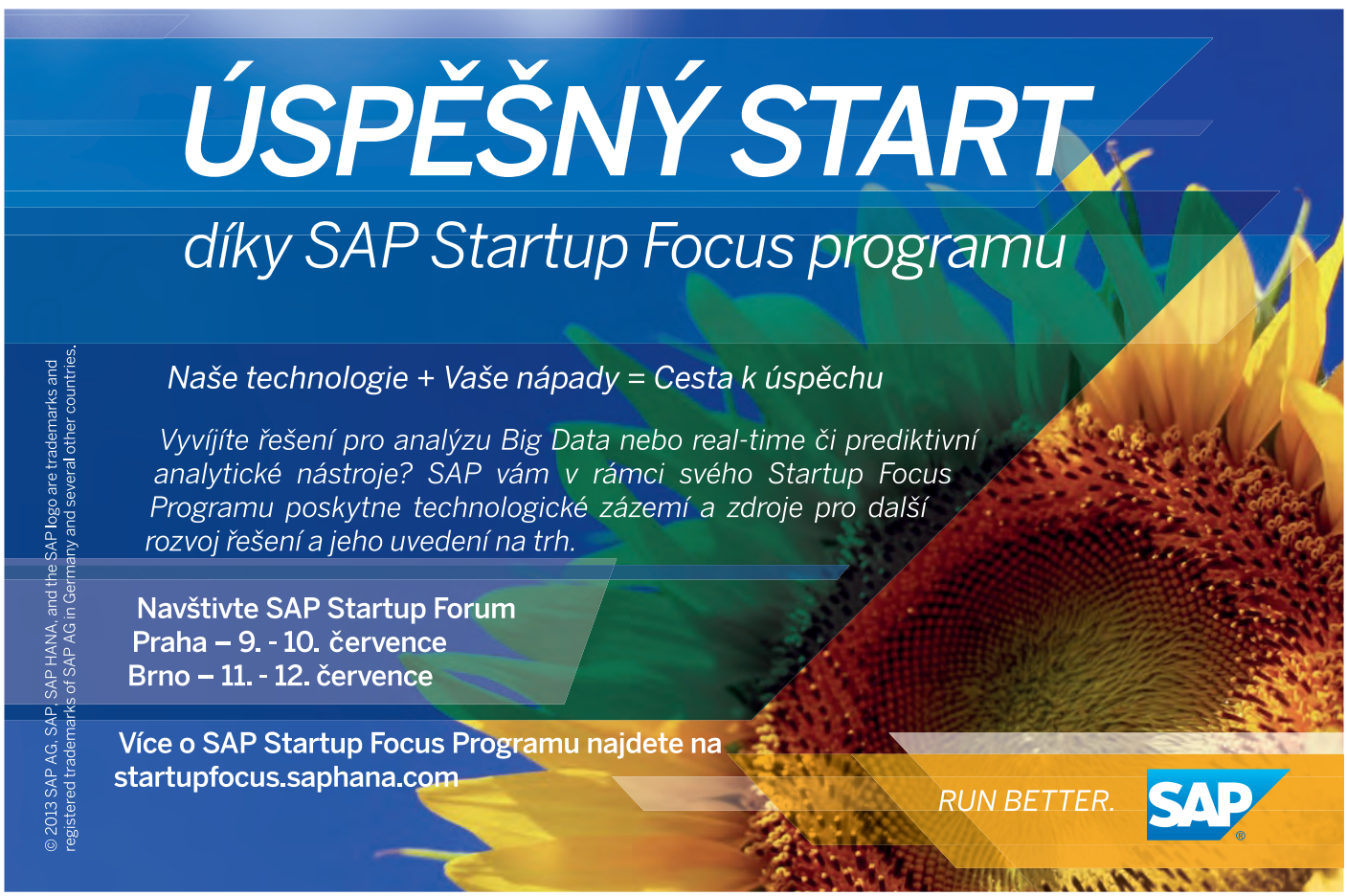
Sada SAP Business Suite na platformě SAP HANA dává firmám dosud nedostupnou možnost analyzovat nejaktuálnější data a na jejich základě činit potřebná obchodní rozhodnutí. Jednotná platforma přitom rovněž snižuje složitost IT prostředí a omezuje redundanci dat i systémů. Podniky tak mohou své procesy a obchodní modely přehodnotit, protože mají k dispozici nástroj, který jim umožňuje to, co dříve nebylo možné.

Sada SAP Business Suite na platformě SAP HANA umožní firmám zrychlit jejich podnikání i zavádění inovací, a otevírá tak zcela nový svět růstových příležitostí.

Firmy mohou nyní v reálném čase řídit všechny své kritické podnikové procesy, jako jsou plánování, realizace, reporting a analýzy, to vše s použitím stejných a aktuálních dat.

Ať už se jedná o CRM systém, ERP nebo analytické řešení, SAP HANA představuje technologii, která až 100 000× urychlí práci s daty a přináší zcela novou třídu analytických funkcí. Přesvědčily se o tom i takové organizace, jako jsou Red Bull, Lenovo nebo Charité a také řada dalších.

Pro toho, kdo chce získat konkurenční výhodu a koho technologie SAP HANA láká, ale přitom stále váhá, existuje v České republice hned několik míst, kde si může řešení postavená na této platformě vyzkoušet, a to třeba i na vlastních datech. Otestovat si lze například ERP nebo CRM systém využívající výhody SAP HANA, malé a střední firmy si zase mohou vyzkoušet SAP Business ONE.


ÚSPĚŠNÝ START
díky SAP Startup Focus programu
Naše technologie + Vaše nápady = Cesta k úspěchu
Vyvíjíte řešení pro analýzu Big Data nebo real-time či prediktivní analytické nástroje? SAP vám v rámci svého Startup Focus Programu poskytne technologické zázemí a zdroje pro další rozvoj řešení a jeho uvedení na trh.
Navštivte SAP Startup Forum
Praha – 9. - 10. července
Brno – 11. - 12. července
Více o SAP Startup Focus Programu najdete na
startupfocus.saphana.com
RUN BETTER.
SAP
© 2013 SAP AG. SAP, SAP HANA, and the SAP logo are trademarks and registered trademarks of SAP AG in Germany and several other countries.

Budoucnost BI je v rychlém zpracování dat

Velké nadnárodní, ale dnes i střední či malé firmy sbírají mnoho dat, která následně využívají např. pro tvorbu finančních a marketingových plánů nebo z nich čerpají nové poznatky pro řízení svého provozu či optimalizaci obsluhy zákazníka.

JAROSLAV PULPÁN

S rostoucím objemem dat se však často informace nedostává včas a ve vhodné formě ke všem příjemcům. Flexibilní, ale přitom jednoduché zpracování do jednoho přehledného rozhraní, které je snem každého manažera, je proto jasným trendem v oblasti business intelligence. Aktuální firemní reporty a analýzy se prostě k uživatelům musí dostávat stejně snadno jako elektronické verze denního tisku.

„Nástroj Visual Analytics od společnosti SAS Institute proto prezentuje informace v dynamickém, několika prsty ovládaném rozhraní na tabletech,“ říká senior konzultant společnosti SAS Institute ČR Jaroslav Pulpán.

Nakolik je SAS Visual Analytics intuitivní pro uživatele?
Tabletem inspirovaná filozofie říká, že uživatelské rozhraní má být navrženo tak, aby nemohlo být jednodušší. Přemísťováním ikon na ploše lze definovat tabulky, analýzy a grafy společně s jejich vazbou na data. Samozřejmostí je automatický výběr vhodných typů grafů či tabulek podle analyzovaných dat. Vzhledy reportů jsou přednastaveny v šablonách, které je možné měnit třeba podle korporátní identity podniku.

Nabízíte tedy řešení, které by mohlo predikovat budoucnost vývoje firmy?
Pomůžeme spíše lépe číst a interpretovat data pro důležitá rozhodnutí managementu s tím, že ukážeme firmám jejich aktivity v úplně jiném světle. Jsme zvyklí pracovat s tabulkami a čísly. Pokud se stejná data zobrazí graficky nebo se například promítnou do mapy, nabízí se zcela jiné možnosti. Firma má najednou velmi detailní a jasný přehled – ví, kolik čeho prodala, na jakém území a kde jsou možné rezervy. Zpracování analýz již netrvá dny či týdny, ale pouhé vteřiny. A to je obrovský krok vpřed pro rychlé a správné rozhodování.

Jak se připravují data pro SAS Visual Analytics?
Smyslem SAS Visual Analytics je zpřístupnit informace koncovým uživatelům s minimálními náklady na IT. SAS Visual Analytics obsahuje administrátorský modul, kterým řídí načítání dat přímo do paměti včetně jejich pravidelné aktualizace. Na rozdíl od dnes tradičních postupů nejsou data nijak agregována. Uživatele tak neomezují předdefinované pohledy.

Administrace dále obsahuje přidělování práv uživatelům a monitoring mobilních zařízení.

S jakými mobilními zařízeními je Visual Analytics kompatibilní?
SAS Visual Analytics pracuje s iPady i tablety s operačním systémem Android. Pro oba typy přístrojů byla vyvinuta mobilní aplikace SAS Mobile BI. Její úvodní stránka zobrazuje ikony nově přidaných analýz či indikuje změny v reportech. Uživatel může na plochu zvolit oblíbené reporty, komentáře a další prvky, v nichž je možné vyhledávat podobně jako v internetovém vyhledávači.

Aplikace dovoluje prohlížení interaktivních výstupů včetně propadu do detailních dat. Všechny prvky reportů a analýz lze on-line komentovat. Komentář je sdílen s ostatními uživateli a lze k němu připojit i další soubory či odkazy podobně jako na sociálních sítích. Díky mobilním zařízením tak manažeři a analytici mohou získat přesný obraz mnohem rychleji než kdy dříve a pomocí mobilní spolupráce rychle reagovat.

Jaká je rychlost zpracování dat ve Visual Analytics?
Protože lidé pracující s tablety nejsou zvyklí čekat, jejich firemní data jsou uložena a zpracovávána přímo v paměti serveru SAS Visual Analytics. To dovoluje uživatelům analyzovat miliony či dokonce miliardy položek. Díky in-memory technologii doba zpracování i v takovémto případě nepřekročí vteřiny.

Pojem big data je v současné době hodně skloňovaný. Existuje hranice mezi normálními a velkými daty?
Hranice určitě neexistuje. Big data je relativní pojem, který má mnoho dimenzí. Gigabajt je pro někoho velký objem dat, pro jiného je jím terabajt, dnes dokonce i petabajt. SAS Visual Analytics ale cílí na uživatele, kteří tyto termíny vůbec nemusí znát. Důležité je totiž podívat se i na jiné aspekty, např. jak často se analyzovaná data mění a jaký je časový horizont, kdy jsou pro mě užitečná.

Gigabajt, který zanalyzovaný vidíte včas ve svém tabletu, má mnohem větší cenu než petabajt v datovém skladu, který prý bude hotov za několik let.

S pojmem big data se hlavně mění smysl ukládání dat. Před deseti lety si společnosti mohly dovolit ukládat objednávky a podle toho velice efektivně přizpůsobovat marketingové nabídky. Big data rozšiřují škálu typů dat, jejichž využití může být profitabilní. Díky poklesu nákladů na sběr dat, zkrácení doby zpracování a metodám prezentace lze uvažovat o dříve nemožných úlohách. Například operativní analýzy nad technickými údaji o provozu telekomunikační sítě nebylo možné ekonomicky ospravedlnit. Dnes lze s takovým objemem dat efektivně pracovat.

Ve které oblasti vidíte potenciál pro zpracování velkých dat?
Ne příliš diskutovaná jsou big bata v oblasti energetiky. Energetická soustava obsahuje desetitisíce měřicích čidel. Blíží se doba, kdy se čidla ve formě elektroměrů rozšíří do domácností. Vzniká zdroj dat využitelný pro různé úlohy optimalizace dodávek energií. Tato bezpochyby velká data lze např. používat pro přesné předpovědi spotřeby či cenotvorbu.

Ale dnes existuje mnoho dalších oblastí, kde pokrok v technologiích, jako je SAS Visual Analytics, úplně změnil poměry v trojúhelníku cena za zpracování – rychlost a forma dodání – dosahované přínosy.

Vidět budoucnost je lepší než jen tušit

Případová studie – Vizualizace big dat v Telecom Italia

Telecom Italia díky analytickým funkcím a vizualizaci dat od společnosti SAS Institute dokáže predikovat klíčové ukazatele a budoucí trendy.

VÍT HRADEČNÝ

Telekomunikační průmysl představuje celosvětově velmi konkurenční prostředí, ve kterém musí společnosti reagovat rychle, aby mohly vycházet vstříc rostoucím a měnícím se požadavkům svých zákazníků. Ti očekávají především rychlost a dostupnost a nejinak je tomu také v Itálii.

Společnost Telecom Italia požadovala flexibilní, uživatelsky příjemné řešení pro vizualizaci velkého objemu dat, která ve společnosti přibývají takřka každou vteřinou. Schopnost rychle zpracovat data a srozumitelně a přehledně je poskytnout i uživatelům, kteří nemají technické či přímo programátorské dovednosti, je alfou a omegou úspěchu firmy.

Díky vizualizaci, která umožňuje inovativní přístup k datům, mohou lidé jednoduše a rychle pracovat. Žádné složité a nepřehledné tabulky. Daleko lépe se pracuje s grafy, předpřipravenými pohledy a šablonami. A hlavně s jednoduchým uživatelským rozhraním, ve kterém lze jednoduše filtrovat, přidávat další pohledy na data, ale také využívat statistické metody, jako je například korelace, jež pomáhá odhalit závislosti mezi daty.

Díky analytickým funkcím je možné predikovat budoucnost a budoucí trendy na základě historických dat. A dělat to rychle a přehledně.

Telecom Italia má zhruba 32 milionů zákazníků a v rámci zlepšování služeb musel rozšířit a posílit schopnost monitorovat síťové služby. Ohromný objem jedinečných a proměnlivých dat však bylo třeba co nejrychleji zpracovat, analyzovat z nich aktuální trendy a činit na jejich základě rychlá a chytrá rozhodnutí.

„Musíme být schopni rychle reagovat a poskytnout našim zákazníkům nové a ještě lepší nabídky. A samozřejmě analyzovat jejich budoucí dopady," říká viceprezident partnerských služeb a vedoucí rozvoje trhu ve společnosti Telecom Italia Fabrizio Bellezza. „Analýza, která má svou cenu a smysl dnes, nemusí již zítra platit. My ovšem potřebujeme vidět mnohem dál než jen zítřek."

Poznání konkurence a nabídka ve správný čas

Aby v Telecom Italia věděli, jak si vedou v porovnání s konkurencí, potřebovali definovat a analyzovat klíčové výkonnostní indikátory (KPI) hlasového a datového provozu mobilních sítí. Na rychle se měnícím trhu, plném zařízení a aplikací, které využívají technologie různých generací, nemusí již zítra platit to, co se dnes považuje za běžné. A v Telecom Italia to potřebují obratem znát. Jen tak totiž mohou porazit konkurenci. Což znamená jediné – vždy mít ve správnou chvíli správnou nabídku pro každého zákazníka.

Jednoduše přístupná data

Využití analytických metod činí z vizualizačních technik řešení, jež umožňuje rychlou a snadnou exploraci dat právě pro uživatele z byznysu. V praxi to znamená, že práce s big daty nevyžaduje, aby se konkrétní pracovníci do hloubky orientovali v problematice analytického zpracování.

Naopak mohou uplatnit svou odbornost ve svých vlastních oborech, jako jsou marketing, finance, zásobování apod. tak, že budou schopni formulovat a zadávat dotazy, které přinesou srozumitelné a okamžitě použitelné odpovědi relevantní pro jejich byznys. Dostávají tak do ruky silný nástroj, jenž jim umožní snadno sledovat a predikovat trendy, rozeznávat závislosti a nepravidelnosti, odhalovat vzorce chování zákazníků, konkurentů, produktů, případně včas zjistit, že některý proces nefunguje dobře. Vizuální forma sdělení je přitom velmi důležitá. Umožňuje snadno pracovat s informacemi vyprodukovanými pomocí tradičních analytických metod, jež v „surovém" stavu nevypadají pro běžné uživatele příliš srozumitelně nebo jsou dokonce „nečitelné".

Nástroj SAS Visual Analytics umožnil společnosti zobrazovat data v uživatelsky příjemném formátu. Díky němu dnes mohou vedoucí pracovníci v Telecom Italia porovnávat výkonnost všech operátorů v oblasti určitého klíčového indikátoru (jako je např. dostupnost nebo podíl přerušených hovorů) na jediné obrazovce, která poskytuje rychlý přehled odpovídajících silných a slabých stránek.

Telecom Italia využívá řešení SAS a jeho schopnosti analyzovat data v paměti a poté je vizualizovat ve spojení s geografickou informací – mapou. To zjednodušuje procesy rozhodování o podpoře i provozu, tak jak se promítají do technického a komerčního plánování.

„SAS Visual Analytics nám pomáhá odhalit nedostatky sítě a zajistit rychle její vylepšení," říká Bellezza. „Umožňuje nám počítat korelace mezi různými KPI a dosáhnout efektivněji hlubší analýzy. A také nám toto řešení umožnilo rozpoznat ziskové oblasti, které můžeme následně posílit z hlediska infrastruktury a služeb poskytovaných na trhu."

Hloubková analýza KPI

SAS Visual Analytics umožňuje firmě Telecom Italia analyzovat celou škálu KPI na různých úrovních agregace jak pro hlasový, tak i datový provoz. Sem patří dostupnost, poměr výpadků, doba vytvoření spojení i datová propustnost, což lze vše sledovat na jediné obrazovce.

„Tím získáváme rychlý přehled oblastí se silnou i slabou konkurenceschopností," vysvětluje Bellezza. SAS Visual Analytics umožňuje společnosti analyzovat pokrytí konkrétních oblastí a identifikovat možné scénáře s prioritou podle nákladů a přínosů. Pomáhá také analyzovat chování zákazníků a vytvářet prediktivní modely, předpovídat služby a hodnotit ziskovost určité oblasti rozvoje po investicích.

„Při prvotním analyzování dat nelze předpovědět, na co se uživatelé mohou dotazovat – a často to nevědí ani sami uživatelé," říká Bellezza. „Nástroj SAS Visual Analytics nám pomáhá porozumět datům, protože zjednodušuje jejich transformace a dovoluje nám převést je na uživatelsky příjemný formát. V důsledku toho vedoucí pracovníci lépe rozumějí tomu, co se děje na trhu," doplňuje Bellezza.

U big dat nestačí jen nakoupit podstatně větší úložiště

Komplexní potřeby a nepřetržité požadavky na zvýšení kapacity trápí správce úložišť. Zde je návod, jak zvládnout záplavu dat.

STACY COLLETTOVÁ

Bývala to kdysi záležitost vědců, internetových gigantů a velkých sociálních sítí, jako Amazon, Twitter, Facebook, Shutterfly. V současné době však stále více podniků všeho druhu chce získat konkurenční výhodu využitím technologií big dat v naději, že se jim odkryjí cenné informace, které mohou různě dostupné údaje obsahovat.

Například společnosti jako Walmart, Campbell Soup, Pfizer, Merck a samoobslužný řetězec Wawa mají s implementací projektů big dat opravdu velké plány.

Někteří se pouštějí do analýz big dat, aby dokázali lépe sledovat informace o zákaznících a okamžitě na ně reagovat nebo aby dostali nové produkty rychleji na trh.

„Každá firma stojí v tomto internetovém věku před faktem, že pokud to neudělá ona, udělá to její konkurence,“ upozorňuje Ashish Nadkarni, analytik segmentu ukládání dat ve společnosti IDC.

Firmy všech velikostí zaplavují data z rozličných interních i externích zdrojů. Mnohé z těchto údajů přitékají v reálném čase a velká část z nich zastará za pouhé minuty, hodiny nebo za pár dnů.

Výsledný nárůst požadavků na ukládání dat je obzvláště znepokojující pro velké podniky, kde si množství strukturovaných a nestrukturovaných dat vyžádalo v průměru o 44 % více úložného prostoru při srovnání let 2010 a 2011, jak uvádí ve své nedávné zprávě agentura Aberdeen Group.

V organizacích všech velikostí se podle analytiků nároky na ukládání dat v průměru zdvojnásobí za každých 2,5 roku. Navíc je pro optimalizaci ukládání videa, tabulek, formátovaných databází a zcela nestrukturovaných dat potřeba nasadit různé dodatečné nástroje.

„Výzvou je pokusit se udržet výdaje na úložiště tak, aby nerostly lineárně se vzrůstajícími požadavky vaší firmy na ukládání dat,“ vysvětluje Dick Csaplar, analytik pro virtualizaci a úložiště ve společnosti Aberdeen Group.

Mezi technologie, které mohou pomoci hlavním uživatelům big dat zabránit tomuto osudu, patří virtualizace úložišť, deduplikace a tiering (vrstvení) úložišť. Pro významné subjekty, jako jsou servery vědeckých institucí a vývojářů simulací či weby sociálních sítí, jsou nejlepší volbou objektová a relačně databázová úložiště.

Praktická stránka systémů navržených k ukládání petabajtů (a mnohdy také více) dat ve snadno dostupném formátu je však mnohem složitější, než je vnitřní fungování běžných platforem pro ukládání dat.

Přinášíme několik rad, jak si správně poradit při správě a ukládání skutečně velkých objemů údajů.

Jaký druh dat analyzujete?

Typ požadovaného úložiště závisí na druhu a množství dat, která analyzujete. Všechny údaje totiž mají omezenou trvanlivost.

Data o akciích jsou například relevantní jen minutu nebo dvě, než se jejich ceny změní. Po výsledku sportovního utkání bývá poptávka 24 hodin nebo do dalšího zápasu.

Tento typ dat je proto potřeba uchovávat v primárním úložišti, když je nejžádanější, a potom ho lze přesunout na levnější úložiště.

Pohled na trendy za poslední roky podporuje myšlenku, že data uložená na dlouhou dobu nemusí být na snadno přístupných primárních jednotkách, aniž to pro jejich majitele představuje nějaký zásadnější problém.

Jak velké úložiště opravdu potřebujete?

Velikost a typ úložiště nutného pro provozování a správu systémů big dat závisí jak na množství ukládaných údajů, tak i na době, po kterou zůstanou tato surová data užitečná.

Existují tři druhy údajů používaných v analýzách big dat, tvrdí Nadkarni.

„Mohou to být data proudící z více zdrojů, zasílaná doslova každou sekundu, a váš čas na zpracování může být jen několik minut, než dojde k jejich zastarání,“ podotýká Nadkarni.

Tento druh údajů obsahuje aktuální informace o počasí, dopravě, sledování trendů témat ze sociálních sítí nebo tweety o událostech po celém světě.

Big data také zahrnují statistické údaje a data vytvářená a kontrolovaná podnikem pro běžný způsob použití.

Data přicházející datovými proudy ale vyžadují jen rychlé zachycení a schopnosti bleskové analýzy, popisuje Nadkarni. *„Jakmile je analyzujete, už je nikdy dále nebudete potřebovat.“*

U statických dat nebo dat vyplývajících z činnosti vaší firmy je tomu ale naopak – musíte je ukládat, vysvětluje Nadkarni.

Jaký typ nástrojů pro ukládání funguje nejlépe?

Podnikům teprve zahajujícím zápas s ukládáním a analýzou big dat doporučují oboroví pozorovatelé virtualizovat svá úložiště, aby bylo vše pod příslovečnou jednou střechou. Dále je vhodné deduplikovat údaje kvůli jejich komprimaci a také použít tiering úložiště, aby se nejcennější data uchovávala na nejsnáze přístupných systémech.

Virtualizace úložiště poskytuje abstraktní vrstvu softwaru, který skrývá fyzická zařízení před zraky běžných uživatelů a umožňuje spravovat všechna řešení jako jeden pool.

Zatímco virtualizace serverů je v současnosti už poměrně dobře zavedenou součástí moderních infrastruktur IT, virtualizace úložišť má ještě co dohánět.

V únoru loňského roku uvedla v průzkumu společnosti Aberdeen, který se uskutečnil mezi 106 velkými organizacemi, jen pětina respondentů, že mají jedinou aplikaci pro správu úložišť. Průměr byl tři aplikace pro správu vztaženou na v průměru 3,2 úložných zařízení.

Mnoho dodavatelů storage systémů se však zdráhá umožnit administraci svých zařízení produktem, který nabízí jiný výrobce. Virtualizace úložišť je totiž *„mnohem složitější a zabere více času, takže zatím nedošlo k takovému přijetí mezi organizacemi, jako je tomu v případě běžné virtualizace serverů,“* prohlašuje Csaplar.

Namísto toho přemýšlí mnoho správců úložišť o cloudových implementacích úložiště takzvané třetí či dokonce čtvrté vrstvy, aby jejich pomocí byli schopni snadněji přesouvat data mezi různými infrastrukturami a snížili své náklady na uchovávání informací.

„Některé společnosti to už udělaly a dosáhly dobrých výsledků, ale není to garantovatelné,“ dodává Csaplar.

Ten v blízké budoucnosti očekává nárůst využití cloudových úložišť a dalších cloudových výpočetních zdrojů v návaznosti na neustálé zlepšování síťové konektivity, snižování cen a zvyšující se schopnosti šifrovat a dešifrovat data bezprostředně při jejich přenosech.

„S cloudem získáte měsíční platby, jež lze hradit z provozního rozpočtu, a ne další položku do investičního budgetu,“ podotýká Csaplar.

Deduplikace a komprese

Správci mohou zmenšit potřebné množství úložného místa prostřednictvím deduplikace, jež eliminuje redundantní údaje pomocí různých datových kompresních nástrojů.

Ty totiž dokážou identifikovat krátké opakující se řetězce v jednotlivých souborech a ve finále uloží jen jednu kopii z nich a na původní pozice umístí jen příslušný odkaz.

Jak velkého snížení požadavků na úložiště lze tímto způsobem dosáhnout? V průzkumu Aberdeenu 13 % respondentů uvedlo, že po nasazení deduplikačních metod snížili objem dat o polovinu, většina podniků však podle analytiků může

Inzerce

IT PRODUKT 2013
COMPUTERWORLD

REDAKCE ČASOPISU COMPUTERWORLD VYHLAŠUJE 3. KOLO SOUTĚŽE

IT PRODUKT 2013

CÍLEM JE VYZDVIHNOUT PRODUKTY S TAKOVÝMI VLASTNOSTMI, KTERÉ PŘINÁŠEJÍ VÝZNAMNÉ POZITIVNÍ ODLIŠENÍ OD KONKURENČNÍCH PRODUKTŮ STEJNÉ KATEGORIE. MŮŽE PŘITOM JÍT O CELKOVĚ INOVATIVNÍ POJETÍ PRODUKTU, POZORUHODNÉ FUNKČNÍ ZDOKONALENÍ, VÝRAZNĚ ZJEDNODUŠENÉ OVLÁDÁNÍ NEBO TŘEBA O VÝJIMEČNĚ PŘÍZNIVOU CENU.

PODROBNÁ PRAVIDLA SOUTĚŽE I DALŠÍ DOPLŇUJÍCÍ INFORMACE NALEZNETE NA ADRESE

CW.CZ/IT-PRODUKT

UZÁVĚRKA 3. KOLA PŘIHLÁŠEK JE 13. 9. 2013

O MOŽNOSTI PŘIHLÁSIT PRODUKTY UVEDENÉ NA TRH PO TOMTO DATU SE PROSÍM INFORMUJTE NA ITPRODUKT@IDG.CZ

počítat spíše se snížením jen o třetinu až polovinu, a to u strukturovaných, vysoce se opakujících dat, poznamenává Csaplar.

Tiering úložišť

Jakmile se podnik rozhodne, jaká data chce analyzovat, mohou administrátoři úložných řešení stanovit ukládání nejnovějších a nejdůležitějších dat na nejrychlejší a nejspolehlivější úložná média.

S tím, jak data stárnou, se mohou přesouvat na pomalejší a levnější úložiště. Systémy automatizující proces tieringu začínají získávat velkou pozornost organizací, ale podle expertů se tyto postupy stále ještě široce nepoužívají.

Při vytváření různých úrovní úložišť musí správci vzít v úvahu technologii konkrétního storage systému, rychlost a propustnost zařízení a také formu RAID, jež je důležitá při ochraně dat před ztrátou a výpadky.

Standardní řešení pro převzetí služeb při selhání je replikace, obvykle v podobě polí RAID.

„Při masivních nasazeních však mohou pole RAID způsobit více problémů, než jich ve skutečnosti vyřeší,“ varuje Neil Day, viceprezident a technologický ředitel společnosti Shutterfly, která provozuje on-line fotoweb umožňující uživatelům ukládat neomezený počet obrázků v původním rozlišení. Úložiště tohoto provozovatele už překročilo kapacitu 30 PB.

V tradičním schématu RAID pro ukládání dat se totiž kopie každé části dat zrcadlí a ukládají na různé disky pole, což zajišťuje jejich integritu a dostupnost. To ale znamená, že uložená a zrcadlená data se mohou nafouknout a v úložišti pak představovat dokonce násobek své velikosti.

Jak se disky používané v polích RAID zvětšují – velmi atraktivní jsou z pohledu hustoty a spotřeby energie 3TB disky –, prodlužuje se také čas potřebný pro obnovu plné parity při výměně vadného disku.

Shutterfly nakonec nasadila technologii samoopravného kódu (erasure code), která rozdělí části dat na kousky, které nejdou samy o sobě použít, a rozptýlí je na různé disky a servery. Data přitom lze kdykoli znovu plně sestavit z jejich částí, přestože jich bylo více ztraceno například vinou poruch disků.

Jinými slovy nemusíte vytvářet více kopií dat – jedna instance může zajistit integritu dat a dostupnost. Vzhledem k tomu, že jsou samoopravné kódy založené na softwaru, lze tuto technologii použít i s komoditním hardwarem, což ještě více snižuje náklady na škálování.

Jedním z prvních dodavatelů softwaru se samoopravným kódem je Cleversafe, který vytvořil to, co nazývá kódování rozptýlením – umožňuje uživatelům ukládat

části v geograficky oddělených místech, jako je více datových center.

Uživatelé obrovských kapacit big dat

Stejně jako Shutterfly musí i další podniky s obrovskými požadavky na úložiště hledat více než jen blokové úložiště, myslí si Nadkarni.

„Když uvažujete o obrovských množinách dat v řádu petabajtů, musíte přemýšlet o objektově založených systémech nebo o distribuovaném souborovém systému,“ prohlašuje Nadkarni.

„Zvažujte komerční nabídky, jako je například škálovatelné úložiště EMC Isilon nebo Dell Fluid File System… a také řešení open source. Jsou mnohem levnější pro ukládání dat a z hlediska výkonu mohou nabídnout mnohem lepší poměr cena/výkon. A v neposlední řadě je lze výborně škálovat,“ dodává Nadkarni.

Uživatelé komerčního softwaru disponují často daty, která jsou částečně jednorázová nebo mají velmi nízké požadavky na následné zpracování, říká Nadkarni.

Menší počet správců

Při správném nasazení by měly virtualizace úložiště, deduplikace, tiering úložiště a samoopravné technologie snížit potřebné množství administrátorů, protože tyto nástroje umožňují řízení prostřednictvím jediného rozhraní.

V případě firmy Shutterfly umožnila automatizovaná úložná infrastruktura zpomalit růst počtu členů týmu údržby. Jakmile v podniku dojde k poklesu nároků na denní údržbu, mohou začít správci trávit více času nad úkoly, které mají pro organizaci vyšší přínos.

„V některých případech projekty big dat uskutečňují speciální týmy a nikoliv tradiční IT personál,“ uvádí Nadkarni. *„Vlastní je a provozují přímo samy obchodní jednotky, protože IT infrastruktura není pro podporu prostředí big dat dostatečně agilní nebo pracovníci IT nemusí mít potřebné znalosti a dovednosti.“*

Podle Nadkarniho dokonce může dojít k situaci, kdy se vůbec neangažují správci úložišť nebo tito administrátoři mohou mít jen relativně malou úlohu, protože v podstatě jen provozovatelům projektu poskytují nějaké úložiště a vše ostatní dělá systémový personál.

Blízká budoucnost

Nadkarni podle svých slov pozoruje trend, při kterém se přesouvá výpočetní vrstva přímo k datům.

„Podívejte se na řešení od společnosti Cleversafe nebo jiných poskytovatelů storage systémů, kteří výpočetní funkce vestavují do samotného úložiště,“ upozorňuje Nadkarni.

„Už nejde zvládnout přesun dat směrem k výpočetní vrstvě. Je to už kvůli stále se zvyšujícímu objemu zpracovávaných informací prakticky nemožné, zvláště pokud máte na analýzu dat jen několik minut, než se stanou nepoužitelná. Proč tedy neumístit výpočetní vrstvu rovnou do místa, kde jsou data?“ ptá se Nadkarni.

Cleversafe nabízí podle Nadkarniho opravdu špičkové řešení založené na Hadoopu a určené pro velké subjekty, jako je třeba zmiňované Shutterfly, ale snaží se i o větší univerzálnost řešení.

„Cleversafe prolamuje model spočívající v pořízení výpočetního výkonu od jednoho dodavatele a úložiště od jiného výrobce,“ tvrdí Nadkarni. Obchodní jednotky podle něj budou muset začít přemýšlet jiným způsobem.

„Jsem si jistý, že se to nakonec ujme, protože současný model prostě nedokáže poskytnout pro big data dostatečnou účinnost,“ říká Nadkarni a dodává: *„Big data představují způsob, jak si udržet svou konkurenční výhodu. Aby společnosti dokázaly ze svých dat získat co největší užitek, budou muset změnit procesy a způsob fungování sebe sama jako firmy a budou také nucené z těchto dat rychle získávat pro sebe klíčovou hodnotu.“*

Dříve, než se lidé začnou zabývat infrastrukturou úložišť big dat, měli by se nad celým problémem velmi vážně zamyslet a důkladně se připravit, myslí si Csaplar.

„Měli by prozkoumat tuto oblast a promluvit si s někým, kdo už to dříve zvládl. Není to už absolutní novinka, takže si můžete promluvit s někým, kdo už má celý proces za sebou, abyste případně neudělali stejné chyby. Na ty už není moc času,“ uzavírá Csaplar. ■

Když Hadoop už nestačí...

Platforma Hadoop podnikům, které potřebují nové a rychlejší způsoby, jak získat obchodní hodnotu z velkých objemů dat, už nepostačuje. Jaké jsou tedy možnosti, díky kterým se bude práce s big daty ještě více akcelerovat?

JAIKUMAR VIJAYAN

Hadoop MapReduce byl dlouho pilířem fenoménu zvaného big data, ale některé firmy v současnosti už potřebují nové a rychlejší způsoby, jak získat obchodní hodnotu z velkých – a stále rostoucích – objemů dat.

Zatímco mnoho velkých organizací stále volí pro své projekty big dat open source prostředí Hadoop, jeho tvůrce, společnost Google, ale i další, už přecházejí na novější technologie.

Apache Hadoop představuje ve své podstatě open source verzi souborového systému Google File System a technologie Google MapReduce. Gigant známý především na poli internetového vyhledávání ji vyvinul pro správu a zpracování obrovských objemů dat umístěných na komoditním hardwaru.

Je hlavní částí zpracovatelských technologií používaných Googlem k procházení a indexování internetu. Hadoop nasadily v průběhu cca uplynulých tří let stovky podniků, aby pomocí něj zvládly rychle rostoucí objem strukturovaných, částečně strukturovaných a nestrukturovaných dat.

Tato open source technologie se ukázala jako levnější alternativa tradiční podnikové technologie pro datové sklady, a to pro účely, jako jsou analýza protokolů a událostí, řízení událostí spojených se zabezpečením, analýzy sociálních sítí a další aplikace zahrnující zpracovávání datových objemů v řádu petabajtů.

Analytici ale poznamenávají, že některé podniky začaly hledat jiná řešení než Hadoop nikoli kvůli tomu, že by tato technologie měla nějaká omezení, ale kvůli účelu, pro který ji její tvůrci navrhli.

Platforma Hadoop je totiž určená především k dávkovému zpracování úloh, kde se údaje shromažďují a zpracovávají v určitých sekvencích. Data se v prostředí Hadoop rozdělují a ukládají v clusteru vysoce distribuovaných komoditních serverů nebo uzlů.

Pokud bude třeba získat z dat nějakou informaci, uživatelé musí nejprve napsat úlohu, odeslat ji a posléze čekat do doby, než se rozdistribuuje na všechny uzly a bude možné zahájit zpracovávání.

Hadoopu chybí rychlost

Ačkoli Hadoop funguje v mnoha případech velmi dobře, není pro některé klíčové aplikace dostatečně rychlý, myslí si Curt Monash, expert na databáze či analýzy a šéf společnosti Monash Research.

Hadoop podle něj například nezvládá dobře interaktivní provoz a ad hoc dotazy pro velké množiny dat. „*Hadoop má potíže s interaktivními odezvami*," zmiňuje Monash a dodává: „*Pokud můžete tolerovat latenci v řádu sekund, potom pro vaše účely Hadoop postačí. Hadoop MapReduce však*

Někteří dodavatelé řešení big dat v České republice

DODAVATEL	OBCHODNÍ JMÉNO	PLATFORMA/SPRÁVA DAT	HARDWARE/ÚLOŽIŠTĚ	ANALYTICKÉ FUNKCE
EMC Czech Republic	Isilon OneFS OS, Isilon Scale-Out NAS, EMC Syncplicity file sync, Greenplum HD (Hadoop Distribution)	proprietární (Isilon, Pivotal One, Pivotal Chorus a Pivotal Analytics) + řešení pro Hadoop (Greenplum HD, Pivotal HD)	Isilon Scale-Out NAS (18 TB AŽ 15 PB), propustnost až 100 GB/s; Isilon Performance Accelerators 10GbE modely (ECC paměť - 8 GB nebo 32 GB, Intel Quad-Core 2.3 GHz X Two)	Pivotal Analytics (Online Business Analytics, IT Operational Analytics a Hadoop Business Analytics)
Hewlett-Packard	HP Vertica	nezávislé na HW a SW platformě, podporuje ANSI SQL-99 +Analytics, R-language, grid computing vč. autom. distribuce zátěže, konektory na Hadoop a Autonomy IDOL10	Dostupný na běžných Unix platformách, DB konektory: JDBC, ODBC, ADO.NET, kapacita neomezena, komprimace dat až 10×, odbavování dotazů až 1 000× rychleji oproti relační DB	Vestavěné funkce: Native gap filling, Interpolace, Event window, Graph traversal, Sessionization
IBM Česká republika	IBM Big Data Platform	Hadoop-BigInsight, DWH, Real Time Engine	DWH	Analytické, prediktivní, statistické, vizualizační funkce
Microsoft Česká republika	Microsoft Big Data řešení	Správa velkých dat on-premise i v cloudu, založené na Hadoop	Dvě varianty - 1. Plug & Play HP/DELL HW + SQL Server 2012, 2. Windows Azure	Jednotné dotazy nad strukturovanými i nestrukturovanými daty, známé nástroje
SAP ČR	Platforma pro datový management - SAP Real-Time Data Platform (SAP HANA rozšířená o nástroje pro monitorování, zprávu a analýzu big dat)	Analýza big dat pomocí SAP HANA in-memory RDBMS a databáze Sybase IQ, integrace s Hadoop a databázemi jiných dodavatelů	Komoditní hardware pro Sybase IQ, HW appliance pro SAP HANA	OLAP a OLTP v jedné databázi - SAP HANA, Podpora NoSQL a SQL, analýza strukturovaných a nestrukturovaných dat, prediktivní analýza, přednastavené analytické funkce
SAS Institute ČR	Visual Analytics - vizualizace dat	Red Hat Enterprise, SUSE Linux Enterprise Server Linux x64 bit; aplikační server JBOSS EAP/ uložení dat Hadoop (dodávaný SAS Institute), EMC Greenplum, Teradata	Dvě varianty - 1. Single server, 16 core CPU, min. 256 GB RAM, 600GB disk, 2. server grid, min. 4 blady např. Dell PowerEdge M610, 96 GB RAM, 12× CPU, 300GB disk per CPU	Analytický server LARS, speciálně vytvářený pro rychlé analytické výpočty, jako jsou deskriptivní statistika, korelace, časové řady a další
Silicon Graphics	SGI DataRaptor, SGI Hadoop Solutions	Hadoop	SGI 21× ISS3124-RP2, 300 TB (504 TB); 20× SGI Rackable servers/160 TB nebo 40× SGI Rackable servers/320 TB nebo SGI Rackable servers/Petabytes	HA disaster recovery, ACID transakce, role-based security, textová analytika, kombinace nestrukt. dat s relačními daty, vizualizační podpora (BI interface) atd.
T-Systems Czech Republic	Analytics as a Service	Cloudera Hadoop Cluster na bázi cloudu	Redundantní cloudové úložiště	Podle potřeb zákazníka, například Microstrategy, Splunk, Empolis a podobně
Teradata Česká republika	Teradata Unified Data Architecture	Proprietární řešení Teradata Aster jako discovery platforma Teradata Enterprise Data Warehouse pro zavádění nových aplikací do produkce	Teradata Aster (až několik PB) pro analytickou práci se semistrukturovanými daty, Hadoop - histore dat, Teradata EDW (až 186 PB) - produkční strukturovaná data	SQL MapReduce - Knihovna analytických funkcí Teradata Aster pro práci s big daty formou SQL, SQL H - SQL a MapReduce analytické funkce nad platformou Hadoop

nikdy nebude vhodná pro latence kratší, než je jedna sekunda.“

Firmy, které potřebují využívat takové vlastnosti, tedy musejí hledat řešení svých analytických potřeb big dat mimo platformu Hadoop.

Google proto začal před přibližně pěti lety používat interně vyvinutou technologii označovanou jako Dremel. Slouží k interaktivní analýze obrovského množství dat protokolů vytvářených jeho tisíci servery po celém světě.

Google tvrdí, že Dremel podporuje interaktivní analýzu velmi velkých objemů dat uložených ve sdílených clusterech komoditních strojů.

Tato technologie zvládne dotazy nad tabulkami s bilionem řádků v řádu sekund a její možnosti škálování sahají do tisíců procesorů a petabajtů dat.

Podporovaný jazyk podobný dotazům SQL usnadňuje uživatelům práci s daty a vytváření ad hoc dotazů, tvrdí představitelé Googlu.

Ačkoli konvenční technologie pro správu relačních databází podporují interaktivní dotazování už mnoho let, Dremel podle tvůrců nabízí mnohem větší škálovatelnost a rychlost.

V provozním prostředí společnosti Google využívají Dremel tisíce uživatelů pro celou řadu účelů – jako jsou například analýza navštívených webových dokumentů, sledování údajů o instalacích aplikací Android, reporty chyb nebo administrace vstupně/výstupních statistik statisíců disků.

Dremel však rozhodně není náhradou za MapReduce a Hadoop, prohlašuje Ju-kay Kwek, produktový manažer hostingové analytické služby big dat založené právě na technologii Dremel. Toto řešení Google nedávno spustil pod obchodním jménem BigQuery.

„Google totiž Dremel používá ve spojení s MapReduce,“ tvrdí Kwek a dodává: *„Hadoop MapReduce se používá k přípravě, čištění, transformaci a uspořádání obrovského množství dat protokolu serveru, přičemž Dremel se následně nasadí pro podrobný rozbor těchto údajů.*

„Hadoop a Dremel jsou distribuované výpočetní technologie, ale každá z nich se vytvořila pro řešení velmi odlišných problémů,“ pokračuje Kwek.

Pokud se například Google snažil vyřešit problém se službou Gmail, bylo třeba prohlédnout obrovské objemy dat protokolu, aby se rychle zjistila příčina.

„Gmail má v současnosti téměř půl miliardy uživatelů. Pokud by měl každý uživatel několik set interakcí s touto službou, jednoduše si můžete představit počet událostí a interakcí, které bychom museli zaznamenávat do protokolů,“ vysvětluje Kwek.

„Dremel nám umožňuje jít do systému a začít zkoumat protokoly pomocí jednoduše položených dotazů,“ uvádí Kwek. Například inženýr Googlu může zadat dotaz: „Ukaž mi všechny odezvy, které byly delší než 10 sekund. Nyní mi je ukaž podle regionu,“ popisuje Kwek.

Dremel podle něj umožňuje inženýrům velmi rychle určit, kde se ono zpomalení ve skutečnosti vyskytlo.

„Technologie Dremelu distribuuje data do mnoha strojů, distribuuje také dotaz na všechny servery a každého z nich se souběžně ptá: ‚máš odpověď na mou otázku?‘ Výsledek pak sloučí a odpověď z obrovského množství zdrojů získá během několika sekund,“ vysvětluje Kwek.

Použití Hadoopu a MapReduce pro stejný úkol by přitom podle Kweka trvalo déle, protože to vyžaduje zapsat úlohu, spustit ji a čekat, až se rozšíří do clusteru – teprve potom se může informace zaslat zpět k původnímu tazateli.

„Můžete to takto udělat, ale je to poměrně chaotické. Je to jako snažit se použít hrníček ke krájení chleba,“ přirovnává Kwek.

Stejný druh objemných dat, který Google před několika lety přiměl k přechodu na Dremel, se začal objevovat také v některých konvenčních podnicích, prohlašuje Kwek.

Změní Google i analytický trh?

Firmy například v automobilovém nebo farmaceutickém průmyslu či v oborech logistiky a financí neustále zaplavují data, a proto jejich IT manažeři hledají nástroje, které jim pomohou rychle je prohledat a analyzovat.

Službu BigQuery vytvořili její tvůrci tak, aby pomohla potřeby technologií big dat plně uspokojit.

„Ve skutečnosti,“ říká Rita Sallamová, analytička Gartneru, *„by mohla hostingová služba založená na Dremelu zcela změnit hru v segmentu analýz big dat.“*

Tato služba totiž podnikům umožňuje pokládat interaktivní dotazy nad obrovskými množinami dat, aniž musí kupovat drahé základní analytické technologie, tvrdí Sallamová.

Firma podle ní může zkoumat různé typy a rozličné objemy dat a experimento-

vat s nimi za zlomek ceny toho, co by stálo koupit komplexní platformu pro analýzu podnikových dat.

Skutečně pozoruhodným aspektem BigQuery ale není základní technologie, nýbrž její potenciál spočívající v radikálním snížení nákladů na IT ve velkých společnostech, upozorňuje Sallamová.

„BigQuery nabízí mnohem ekonomičtější způsob, jak analyzovat velké objemy dat ve srovnání s tradičními podnikovými platformami dat a má skutečně potenciál snižovat celkové výdaje. Také společnostem umožňuje experimentovat s vlastním obsahem big dat,“ podotýká Sallamová.

Konkurence ke Googlu

Hlavní dodavatelé produktů pro business intelligenci, jako jsou firmy SAS Institute, SAP, Oracle, Teradata nebo Hewlett-Packard, také chtějí dodávat nástroje, které poskytují lepší možnosti analýzy dat.

Stejně jako Google považuje většina těchto výrobců platformu Hadoop především za úložiště velkého množství dat pro následnou přípravu a uspořádání dat k následné analýze jinými nástroji.

Například v říjnu loňského roku SAP představil nový balík pro big data navržený tak, aby velké organizace mohly integrovat prostředí Hadoop s její in-memory databází SAP HANA a se souvisejícími technologiemi.

Balík využívá platformu HANA ke čtení dat z prostředí Hadoop a následně nabízí rychlé interaktivní analýzy údajů pomocí reportovacích a analytických nástrojů SAP.

Firma SAS oznámila podobné funkce pro svůj High Performance Analytic Server ještě o něco dříve. HP s technologií získanou při akvizici firmy Vertica, společnost Teradata s Aster-Hadoop Adaptorem či IBM se sadami nástrojů Netezza rovněž nabídnou podobné schopnosti.

Tento byznys také přilákal pár nových firem. Jedna z nich, Metamarkets, například vyvinula cloudovou službu pomáhající společnostem analyzovat velké množství čerstvých dat přitékajících v reálném čase.

„Jádrem naší služby je interně vyvinutá technologie distribuované in-memory sloupcové databáze, nazývaná Druid,“ uvádí Michael Driscoll, výkonný ředitel společnosti Metamarkets.

Koncepci vlastní technologie Druid přitom přirovnává k výše popsanému řešení Dremel od Googlu.

„Technologii Dremel její tvůrci od začátku navrhli tak, aby fungovala jako analytické úložiště dat,“ popisuje Driscoll a dodává: *„Její sloupcově orientovaná, paralelizovaná in-memory architektura z ní dělá o několik řádů rychlejší řešení, než jsou tradiční storage systémy.“*

„My máme velmi podobnou architekturu,“ tvrdí Driscoll. *„Používáme sloupcovou orientaci, distribuovanost i řešení postavené na technologii in-memory.“*

Systém od firmy Metamarkets však podnikům umožňuje dotazovat se nad daty ještě před tím, než ve skutečnosti dotečou do datového úložiště – tak podle tvůrců nabízí rychlejší vhled, než je tomu v případě Dremelu, zdůrazňuje Driscoll.

Firma Metamarkets uvolnila v loňském roce technologii Druid pro komunitu open source, aby kolem této technologie podnítila větší vývojářskou aktivitu. *„Poptávka po ní se řídí všeobecnou potřebou rychlosti,“* připomíná Driscoll.

Hadoop, jak podotýká Driscoll, je zkrátka příliš pomalý pro organizace, které potřebují u svých dotazů čas odezvy v řádu milisekund.

„Analytické technologie, jako jsou ty, které nabízejí tradiční podnikoví dodavatelé, jsou sice rychlejší než samotný Hadoop, ale stále ještě neposkytují takové škálování, jako je tomu v případě Dremelu nebo Druidu,“ zdůrazňuje Driscoll.

Nodeable je další novou firmou v tomto oboru. Poskytuje cloudovou službu nazývanou StreamReduce, která se nabídce Metamarkets zčásti podobá.

StreamReduce využívá analytickou open source technologii Storm, kterou původně vyvinula firma BackType předtím, než ji v roce 2011 koupil Twitter. Technologie Storm, která se mimo jiné interně využívá právě Twitterem, umožňuje podnikům analyzovat datové proudy v reálném čase.

Nodeable zároveň nabízí propojení pro Hadoop, aby podniky mohly používat službu také ke spouštění interaktivních dotazů nad daty uloženými právě v prostředích Hadoop, uvádí Dave Rosenberg, výkonný ředitel firmy Nodeable.

Ta přitom vznikla původně jako organizace zabývající se správou cloudového systému, ale své zaměření změnila poté, co si povšimla příležitostí spojených s technologiemi analýz big dat.

„Uvědomili jsme si, že tam pro Hadoop chybí doplněk, který by umožnil zpracování v reálném čase. Položili jsme si proto otázku, jak takové podpory s pomocí Hadoopu dosáhnout,“ popisuje Rosenberg.

„Služby jako Nodeable řešení Hadoop tedy nenahrazují, nýbrž ho doplňují,“ vysvětluje Rosenberg.

StreamReduce poskytuje organizacím způsob, jak extrahovat užitečné informace z datových proudů, které se mohou ukládat v prostředí Hadoop nebo v jiném úložišti kvůli pozdějšímu tradičnějšímu dávkovému zpracování, dodává Rosenberg.

Stroje, které zpracovávají datové proudy, jako jsou ty z nabídky společností Nodeable a Metamarkets, se od technologií, jako je Dremel, liší v jednom důležitém aspektu – jsou vhodné pro analýzu surových dat ještě předtím, než se uloží do příslušné databáze.

Dremel a další podobné technologie jsou naopak vhodné pro pokládání ad hoc dotazů nad daty, která jsou už uložena ve storage systému, jako je třeba prostředí Hadoop.

Ani Hadoop nezůstává stranou

Ani další hráči aktivní v současném prostředí platformy Hadoop však mezitím jen nečinně nepřihlížejí. Například Cloudera, největší dodavatel komerční podoby Hadoopu, nedávno představila technologii Cloudera Impala. Ta je ve své podstatě strojem pro dotazy v reálném čase nad daty uloženými v distribuovaném souborovém systému Hadoop.

Technologie Impala umožní firmám dělat operace v reálném čase i dávkové úlohy nad strukturovanými i nestrukturovanými daty v rámci jediného systému, uvádí Cloudera. ■

T · · Systems ·

Máte včas podklady pro kvalifikovaná rozhodnutí?

ZDENĚK LEJSEK

Objem uložených dat neustále roste, tak jak přibývá zdrojů, které je mohou poskytovat. To platí, ať už jde o podnikové informační systémy, které sledují stále více parametrů, data nasbíraná v průběhu výroby či data získaná z nejrůznějších monitorovacích systémů, jako jsou třeba kamerové systémy nebo nově i sociální sítě.

IT obor si dokáže dobře poradit s jejich ukládáním, otázkou ale je, zda uložená data dokážeme smysluplně využít. A tady je problém, protože dosavadní metody pro zpracování velkého objemu dat přestávají stačit. Zde se totiž střetávají dvě protichůdné síly – rychlost, jakou jsme schopni data zpracovávat, a rostoucí požadavky na to, aby výsledky byly k dispozici včas. Co by bylo platné, kdyby systém pro předpověď počasí zpracoval prognózu na následující den s přesností 95 % během 48 hodin.

A podobné je to v podnikání. Pokud chcete reagovat včas na rychlé nebo skryté změny tržního prostředí, musíte dosáhnout toho, aby se doba zpracování zkrátila tak, aby vám výsledek umožnil reagovat na situaci v reálném čase. Je jasné, že společnosti, které dokážou včas zpracovat velké objemy dat na podklady pro kvalifikovaná rozhodnutí, mohou získat významnou tržní výhodu.

První krok, jak si poradit s rostoucím objemem, je obvykle snadný: agregace dat. V řadě případů jde o konsolidaci „pouhých“ stovek terabajtů. A technologie, která může vyřešit tyto otázky, je už chvíli na scéně – cloud computing. Cloud sehrává hlavní úlohu při transformaci velkých objemů dat do informace potřebné pro operativní akci. Je nástrojem pro sledování trendů, které již dlouho představují výzvu pro podniky, IT manažery a pro správu obrovských složitých souborů dat.

Cloud je navíc odrazovým můstkem k řešením, která podporují metriky v reálném čase a poskytují čerstvé nové pohledy – řešení, která využívají předpovědní analýzy k simulaci a predikci chování zákazníků. Firmy schopné zvládat obrovské objemy informací ovládají rozmanité zdroje, řídí rychlost a uznávají hodnoty, nepovažují velké objemy dat za problém, nýbrž je chápou jako řešení.

Autor pracuje ve společnosti T-Systems Czech Republic jako specialista marketingu

Analýza rizik v reálném čase

Byznys: singapurský poskytovatel finančních služeb s 500 pobočkami v 19 zemích

Výzva: analyzovat rizika v reálném čase

Řešení: technologie pro výpočty v paměti a vysokovýkonné analytické nástroje

Přínosy: analýzy rizik nad 8,8 mld. kalkulací rizik zkráceny z 18 hodin na několik minut

United Overseas Bank (UOB) ze Singapuru patří s 237 mld. aktiv k vedoucím poskytovatelům finančních služeb v oblasti jihovýchodní Asie. Během pěti let vyskočily její akcie o 45 %, což dokazuje, jak dobře se vedení společnosti orientuje v proměnlivém světě bankovnictví. Analýza rizik je v bankách velmi komplexní úlohou. V UOB jsou rizika rozložena na 45 tisíc finančních instrumentů a jsou ovlivňována více než stovkou tisíců tržních parametrů včetně cen a dob splatností. Kalkulace celkových rizik banky znamená pro UOB provést kolem 8,8 miliardy vysoce komplexních hodnocení rizik.

Rychlé odpovědi

Až donedávna IT oddělení banky vedené CIO Susan Hweeovou potřebovalo 18 hodin na vyhodnocení vlivu tržních změn na celkové riziko banky. To znemožňovalo získat hodnocení nových rizik ad hoc. Ať už šlo o rozhodnutí k záchraně eura ve vzdáleném Bruselu, přírodní katastrofy nebo nové video s Mohamedem, IT oddělení zjistilo, že politické, sociální a ekonomické události se na trzích neprojeví ve dnech, ale spíše během několika minut.

V případě UOB tak nebyl problém ani v objemu dat, která bylo třeba analyzovat, jako spíš v rychlosti, jakou bylo nutné rizika vyhodnocovat. Aktuální analytický systém banky prostě už nebyl dostatečně rychlý. To bylo důvodem pro to, aby začátkem roku 2012 začalo IT oddělení vyhledávat vhodné nástroje na zpracování big dat.

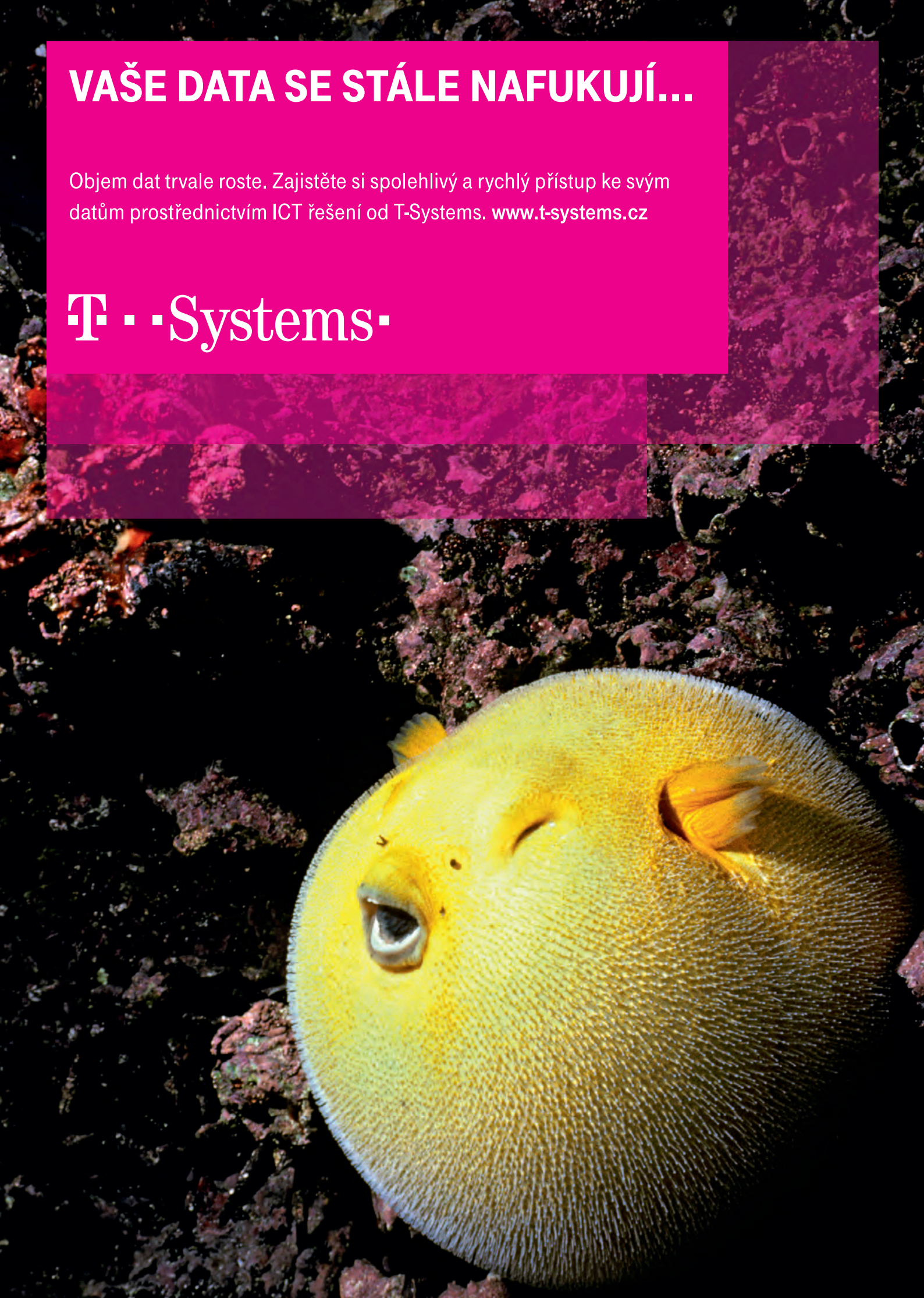
V současné době testované řešení stojí na dvou pilířích: výpočtech v paměti a výkonném analytickém softwaru od společnosti SAS. Za pomoci tohoto systému mohou být rizika vyhodnocena v řádu minut, což znamená, že se analýzy dějí téměř v reálném čase a mohou zohlednit i současné změny parametrů.

V minulosti byla analýza rizik chápána především jako únavný proces vyžadovaný regulačními orgány. Nyní slouží analýza rizik UOB jako obchodní nástroj – k hodnocení strategií a zvážení vlivu trhu a dění ve světě.

VAŠE DATA SE STÁLE NAFUKUJÍ...

Objem dat trvale roste. Zajistěte si spolehlivý a rychlý přístup ke svým datům prostřednictvím ICT řešení od T-Systems. **www.t-systems.cz**

T · · Systems ·



Cesta k efektivnímu zhodnocení podnikových informací

Společnost Teradata se zaměřuje na zpracování „velkých“ dat, ve smyslu analytického využití velkých datových objemů pro Business Intelligenci, už více než 30 let. Unikátní vlastnosti databázové platformy Teradata, jako jsou masivně paralelní zpracování databázových dotazů, lineární škálovatelnost hardwaru a softwaru a dynamické řízení zdrojů při zpracování komplexního databázového workloadu, umožňují platformě Teradata garantovat požadovaná SLA pro řádově vyšší datové objemy než ostatní databázové systémy.

JONÁŠ SVATOŇ

V posledních několika letech se však mění chování společností v oblasti zpracování podnikových dat. Jde především o snahu využít nové datové struktury a datové zdroje, jejich postupné zapojování do podnikové analytiky a využití jak v BI, tak v obslužných procesech zákazníka. V rámci moderních trendů využití nově dostupných dat je zavedena kategorie – big data, která je charakterizována nejen svým rozsahem, ale zároveň i svou semistrukturovanou formou, která není vhodná pro tradiční relační databázové systémy. Big data obecně jsou dále spojená s dosud malou business znalostí toho, jaká část těchto dat může společnostem reálně přinést business hodnotu. Vzhledem k výše uvedeným charakteristikám této datové kategorie je ekonomicky neefektivní historizovat tato data ve standardních datových úložištích, která běžně reprezentuje podnikový datový sklad.

Východiska Teradata Unified Data Architecture

Ukazuje se, že jedna platforma není ekonomicky a technologicky výhodná pro práci se všemi typy dat. V souvislosti se zapojováním nových datových zdrojů do podnikových analytických procesů roste i škála analytických nástrojů, které jsou pro tuto práci využívány. Dosavadní SQL platformy již nedostačují pro pokrytí všech těchto potřeb.

Dlouhodobě preferované paradigma pohledu na jednotné podnikové datové úložiště, které integruje a historizuje data ze všech provozních systémů do businessově orientovaného modelu, doznalo změn, které byly promítnuty v Teradata Unified Data Architecture (UDA). Tato architektura je postavena především na třech základních principech:

- Pro efektivní využití big dat musí být business analytikům v celé organizaci poskytnuty správné analytické nástroje jak pro práci se stávajícími zdroji dat, tak s daty, která jsou nově k dispozici
- Datová architektura musí využívat kombinaci vhodných technologií pro rozdílné analytické úlohy
- Podniková analytická a BI infrastruktura musí být vzájemně propojená tak, aby jednotlivé komponenty mohly datové a informační vstupy sdílet

Jako reakci na typické kategorie business požadavků v rámci zpracování big dat společnost Teradata zavádí tři základní komponenty podnikového datového analytického prostředí, Unified Data Architecture:

- Platforma pro levné uložení velkých objemů dat především nestrukturovaného charakteru. Běžně bývá tato komponenta reprezentována technologií Hadoop.
- Discovery platforma pro zjednodušení práce se semistrukturovanými daty a propojení strukturovaných a nestrukturovaných dat nabízí business analytikům známé nástroje pro exploraci dat, především SQL, a minimalizuje závislost na znalosti sofistikovanějších programovacích jazyků, jako je Java nebo C/C++.
- Platforma integrovaných podnikových dat pro zavádění nových zjištění z Discovery platformy do produkčních procesů vedle provozního reportingu, dataminingového skóringu a dalších BI aplikací.

Technologicky Teradata svými produkty a produkty svých partnerů pokrývá celé portfolio prostředí Unified Data Architecture.

Hadoop pro uložení a předzpracování dat

Mnoho společností již nyní ukládá velké množství semistrukturovaných historických dat na této platformě. Jedná se především o logy návštěvnosti webových stránek, multimediální obsah, data ze senzorů nebo technologická data o fungování telekomunikační sítě. Protipólem levného uložení těchto dat je obtížné získávání nových poznatků a jejich produkční využití. Důvodem je především náročný vývoj a analytická práce přímo nad touto platformou vyžadující specializované programovací techniky.

V rámci UDA Teradata doporučuje těmto společnostem využít stávajících investic pro uložení a předzpracování těchto dat, nicméně přesunout většinu analytické práce do uživatelsky přívětivějšího prostředí Discovery platformy.

Teradata Aster jako Discovery platforma pro semistrukturovaná data

Platforma Teradata Aster umožňuje rychlé nahrání semistrukturovaných dat z Hadoop nebo jiného úložiště a jejich jednoduché převedení do strukturovaného formátu. K této práci Teradata Aster využívá patentovaný framework SQL MapReduce, který zajišťuje například automatickou konverzi webových logů do strukturované formy, na kterou lze následně aplikovat standardní analytické techniky SQL jazyka. Business Analytik tak není závislý na vývoji v Javě a jiných programovacích jazycích při samotném používání Teradata Aster.

Požadavky na big data jsou řešeny kombinací technických komponent analytické databáze, Discovery platformy a non-SQL řešení typu Hadoop

Platforma umožňuje využití desítek předpřipravených funkcí SQL MapReduce včetně jejich vývoje na míru v jazycích Java C/C++, Python či R. Nově vytvořené funkce, spouštěné prostřednictvím SQL dotazů, je pak možné používat pro analýzu semistrukturovaných dat.

Typické analytické funkce Discovery platformy lze zařadit do následujících kategorií:

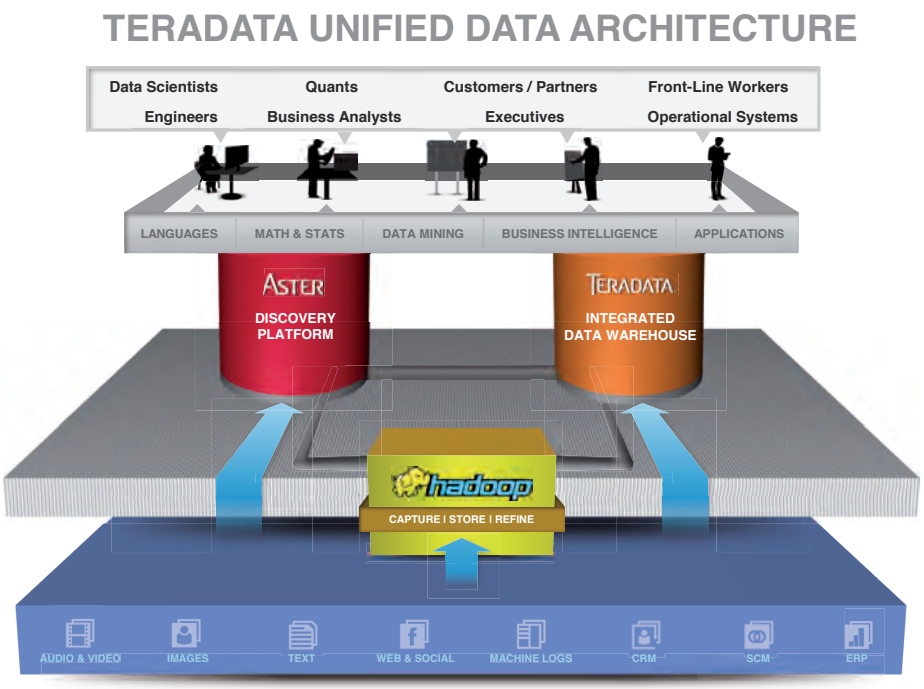
- Analýza řad časově závislých událostí (funkcionalita nPath), které vedly ke zkoumanému výstupu, například nákupu na e-shopu, nebo naopak opuštění košíku
- Transformace semistrukturovaných dat do strukturované formy – například sjednocení všech sessions v rámci webového logu, které představují jednu návštěvu webových stránek
- Analýza vztahů mezi jednotlivými datovými prvky – například pro řešení analytiky nákupního košíku
- Analýza textu – například sentiment analýza vztahu k firemní značce nebo produktu na sociálních médiích

Klíčovým prvkem jsou konektory mezi Discovery platformou a podnikovým datovým skladem, které umožňují analytikům obohatit analýzy semistrukturovaných dat o strukturované informace z datového skladu. Analytik má například možnost doplnit analýzu aktuálních přístupů na webové stránky o historii nákupů jednotlivých zákazníků a poskytnout doporučení produktu na míru konkrétního zákazníka pro jeho příští návštěvu portálu.

Platforma dále umožňuje práci s daty přímo v prostředí Hadoop díky své další funkcionalitě, SQL-H. Tato technika eliminuje nutnost přesunů dat mezi jednotlivými platformami a dává analytikům možnost psaní SQL dotazů spouštěných prostřednictvím databáze Aster přímo nad úložištěm Hadoop s využitím metadat v HCatalogue.

Teradata jako Enterprise Data Warehouse

Teradata EDW primárně obsahuje integrovaná podniková data ve strukturované formě, reprezentované businessově orientovaným logickým datovým modelem.



Společnost Teradata zavádí Unified Data Architecture framework pro ekonomicky efektivní řešení práce s big daty

Nové analytické aplikace, jejichž hodnota je identifikována v Discovery platformě, jsou následně zaváděny do produkčního zpracování na této komponentě.

Podnikový sklad obsahuje komponenty pro pravidelné nahrávání strukturovaných dat z provozních systémů společnosti a zároveň výstupů analýz nad nestrukturovanými daty z Discovery platformy nebo přímo z Hadoop prostředí. Nad těmito daty je postaven především podnikový reporting, produkční úlohy dataminingu a zajištění taktického přístupu k podnikovým informacím z odběrných aplikací, jako jsou Portál, aplikace Call centra nebo CRM.

Přínosy UDA pro zpracování big dat

Teradata Unified Data Architecture přináší především rozšíření spektra dat využitelných v analytice a BI o nové datové typy a zajištění informační hodnoty semistrukturovaných dat pro podporu rozhodování v rámci zákaznických a provozních procesů společnosti. Respektuje principy racionalizace nákladů na IT pro ukládání a využití celého spektra podnikových dat a především ochranu stávajících investic do podnikového datového skladu a do relativně levných úložišť semistrukturovaných dat. Doplnění stávající architektury o Discovery komponentu zjednodušuje práci analytiků, snižuje nároky na jejich znalosti programovacích jazyků nebo jejich závislost na programátorech v IT. Kombinace těchto prvků společnostem budujícím infrastrukturu pro práci s big daty na bázi Teradata Unified Data Architecture přináší především zkrácení time to market výstupů analytických činností.

Autor je Senior Business Consultant společnosti Teradata Česká republika

NoSQL: Ne pro všechno se hodí

Módní termín NoSQL se šíří podnikovým IT už několik let. Nadšení z těchto rychlých datových skladů hojně využívaných právě pro big data bylo opojné a dojem průkopnických organizací z NoSQL velmi lákavý. Líbánky se však chýlí ke konci a je čas začít vyvažovat nadšení i některými poměrně nepříjemnými pravdami.

PETER WAYNER

Nechápejte článek špatně. Stále doporučujeme zkoušet nejnovější varianty řešení pro vytváření jednoduchého mechanismu ukládání dat. Stále si uvědomujeme velkou hodnotu systémů, jako MongoDB, CouchDB, Cassandra, Riak a dalších projektů spojených s NoSQL.

Pořád je dobré svěřovat některá z nejdůležitějších firemních dat těmto technologiím, protože jsou stále lepší a každý den otestované dalšími bitvami.

Musíme si však více uvědomovat také slabiny, protože systémy NoSQL jsou vzdálené dokonalosti a často přinášejí potíže. Nejchytřejší vývojáři to věděli už od začátku. Nespálili proto SQL manuály ani mosty k prodejním kanálům SQL.

Prozíraví IT manažeři NoSQL jednoduše konstatovali, že NoSQL je spíše ‚Not Only SQL', tedy jinými slovy spíše variantní řešení doplňující současné široce rozšířené SQL systémy. Pokud masy chápaly zkratku špatně, byl to jejich problém, na který teď možná doplácejí.

Níže uvedený seznam problémů, ať již velkých nebo malých, je tedy pokusem všechny tyto skutečnosti zdokumentovat a takříkajíc vyčistit vzduch.

Slouží k uvedení věcí do pořádku, abychom dokázali lépe pochopit výhody a nevýhody bezesporu zajímavých systémů NoSQL.

Nepříjemný fakt NoSQL Č. 1

JOIN znamená konzistenci

Jedna z prvních stížností lidí na SQL systémy poukazuje na výpočetní náročnost vykonávání příkazů JOIN mezi dvěma tabulkami. Základní myšlenkou je ukládat data jen na jednom místě.

Pokud tedy udržujete seznam zákazníků, vložíte jejich adresy do jedné tabulky a zákaznické ID použijete v každé další tabulce. Když získáváte data zpět, příkaz JOIN udělá propojení ID s adresami – a vše tak zůstává konzistentní.

Potíž ale nastává v tom, že JOIN může být poměrně náročný, a někteří správci databází dokonce vymysleli tak složité příkazy JOIN, že jsou matoucí a dokážou srazit na kolena i nejrychlejší hardware.

Nebylo tedy překvapením, že vývojáři NoSQL překlopili příkaz JOIN do vlastnosti: Udržujme adresy zákazníků ve stejné tabulce jako všechno ostatní! Cesta NoSQL je ukládat dvojice klíč – hodnota pro každou osobu. Když nadejde čas, načtete je všechny.

Bohužel lidé, kteří chtějí, aby jejich tabulky zůstaly konzistentní, stále potřebují příkaz JOIN. Jakmile začnete ukládat adresy zákazníků se všemi podrobnostmi, často skončíte s více kopiemi těchto adres v každé tabulce. A když máte více kopií, je třeba je aktualizovat všechny. Někdy to funguje, ale pokud se tak nestane, nedokáže to NoSQL řešit pomocí transakcí.

Počkejte, říkáte si, proč nemít samostatnou tabulku s informacemi zákazníka? Potom bude potřeba měnit jen jeden záznam. To je skvělý nápad, ale dostali jste se zpátky k příkazu JOIN.

Nepříjemný fakt NoSQL Č. 2

Složité transakce

Řekněme, že můžete fungovat bez problému, aniž použijete příkazy JOIN nad tabulkami – protože potřebujete rychlost. Je to přijatelný kompromis a někdy správci databází SQL denormalizují tabulky právě z tohoto důvodu.

Potíž nastává v tom, že NoSQL ztěžuje udržení konzistence různých záznamů. Často neexistují žádné transakce, které by zajistily, aby se změny udělaly najednou ve více tabulkách. To musíte zabezpečit sami a chyba by mohla způsobit, že vznikne nekonzistence tabulek.

Rané implementace systémů NoSQL tyto transakce opomíjely. Nabízely seznamy dat tam, kde byly konzistentní, s výjimkou případů, kdy tomu tak nebylo.

Jinými slovy se snažily získat data s nejnižší hodnotou, kde případná existence chyb nezpůsobí žádný významný rozdíl.

Některé současné implementace NoSQL ale nabízejí něco, co se blíží transakci. Například NoSQL produkt od společnosti Oracle umožňuje transakční kontrolu nad daty zapisovanými do jednoho uzlu a výběr flexibilního rozsahu konzistence ve více uzlech.

Pokud chcete perfektní konzistenci, budete muset počkat, než se každý zápis uskuteční ve všech odpovídajících uzlech. S přidáním další struktury a ochrany tohoto typu experimentuje i několik dalších systémů typu NoSQL.

Nepříjemný fakt NoSQL Č. 3

Flexibilita schématu je tikající bombou

Jednou z velkých myšlenek modelu NoSQL je, že nevyžaduje schéma. Jinými slovy, programátoři nemusí dopředu určit, které sloupce mají být k dispozici pro každý řádek v tabulce.

Jedna položka může mít například 20 připojených řetězců, další obsahovat 12 čísel a jiná může být zcela prázdná. Programátoři mohou udělat rozhodnutí, kdykoli potřebují něco uložit. Nemusí tedy žádat o povolení správce databáze a vyplňovat všechny „papíry“, aby mohli přidat nový sloupec.

Všechna tato svoboda zní opojně a ve správných rukou může urychlit vývoj. Je to ale opravdu dobrý nápad pro databázi, která může fungovat se třemi týmy vývojářů? A je to životaschopný model pro databázi, která by mohla existovat déle než šest měsíců?

Jinak řečeno, vývojáři mohou chtít svobodu pro úpravy databáze, ale chtěli byste být pátým vývojářem, jenž přijde po čtyřech předchozích, kteří si volili své vlastní klíče?

Je snadné vymyslet různé reprezentace „narozenin“, které každý vývojář vybere pro klíč, když se bude přidávat k záznamu narození uživatele. Tým vývojářů může vymyslet téměř cokoliv: „bday“, „b-day“, „birthday“ apod.

Struktura NoSQL nenabízí žádnou podporu pro omezení tohoto problému, protože by to znamenalo vnucení schématu. Nechce omezovat skvělé vývojáře, schéma by jim překáželo.

Faktem je, že přidat sloupec do tabulky není velký problém, a určitá kázeň může být ve skutečnosti dobrá i pro vývojáře. Stejně jako je nutí k určení typů proměnných, pobízí je i v tomto případě k tomu, aby určili typ dat připojených ke sloupci.

Ano, správce databáze může přinutit vývojáře k vyplnění formuláře v trojím vyhotovení před připojením sloupce, ale není to tak zlé jako potýkat se s pěti klíči vytvořenými za pochodu.

Nepříjemný fakt NoSQL Č. 4

Databáze mohou být inteligentní

Mnoho programátorů NoSQL se rádo chlubí tím, jak jejich odlehčený kód a jednoduchý mechanismus pracují velmi rychle. Mají obvykle pravdu, když jsou úkoly stejně jednoduché, jako je vnitřek NoSQL, ale ke změně dochází, pokud se úkoly ztíží.

Zamysleme se nad výše popsaným problémem ohledně příkazu JOIN. Jakmile začali programátoři vytvářet své příkazy JOIN podle vlastní logiky, pokoušeli se to udělat efektivně.

Vývojáři SQL strávili desítky let vývojem důmyslných strojů pro zpracování příkazů JOIN, aby pracovaly tak efektivně, jak to jen bylo možné.

Jeden z expertů na SQL dokonce řekl, že se snaží synchronizovat svůj kód s rotujícím pevným diskem, aby vyžadoval údaje jen tehdy, když jsou hlavy na správném místě.

Může se to zdát extrémní, ale SQL vývojáři skutečně pracují na podobných vychytávkách po celá desetiletí.

Není tedy pochyb o tom, že programátoři tráví celé dny horlivým úsilím o dosažení strukturovanosti svých SQL dotazů, aby maximálně využili výhody skryté inteligence. Nemusí být jednoduché to využít, ale když na to programátor přijde, může začít databáze příslovečně „zpívat“.

Sofistikovaný dotazovací jazyk SQL má vždy potenciál zastínit jednoduchý dotazovací jazyk, který je dostupný v řešeních NoSQL.

U prostých případů to nemusí mít význam, ale když se situace stává složitější, SQL dotazování se dělá na stroji přímo u dat. Načítání dat a vykonání práce mají potom malou režii. Server NoSQL musí obvykle zaslat data na místo určení.

Nepříjemný fakt NoSQL Č. 5

Příliš mnoho přístupových modelů

Čistě teoreticky – SQL představuje standardní jazyk. Pokud používáte SQL pro jednu databázi, měli byste být schopni spustit stejný dotaz také v jiné kompatibilní verzi.

Toto tvrzení může fungovat s několika jednoduchými dotazy, ale každý správce databáze ví, že může trvat roky, než se naučí výstřednosti SQL pro různé verze stejné databáze. Dochází ke změně definic klíčových slov a dotazy pracující s jednou variantou nemusí pracovat s jinou.

Technologie NoSQL je ještě v mnohem vyšší míře „jen pro zasvěcené“. Je to jako Babylónská věž. Od začátku se všichni vývojáři NoSQL pokoušeli představit si nejlepší možný jazyk, ale mají pro to velmi odlišné představy.

Toto semeniště experimentů je dobré – dokud se nepokusíte přejít mezi různými nástroji. Dotaz pro CouchDB se vyjadřuje

Inzerce

COMPUTERWORLD na papíru, v počítači, mobilu i tabletu za jednu cenu!

Roční předplatné tištěné a elektronické verze Computerworldu nyní **jen za 429 korun**. Jako **bonus navíc** i roční předplatné tištěné verze **CIO Business World** a **SecurityWorld**.

Přečtěte si svůj Computerworld opravdu kdekoliv.

Za jedno zvýhodněné předplatné nyní získáte Computerworld nejen v tištěné variantě, ale také jeho elektronickou verzi vhodnou pro stolní počítače, mobily i tablety.

Zvýhodněné předplatné objednávejte na adrese **http://cw.idg.cz**

jako dvojice funkcí JavaScriptu pro mapování a redukci.

Rané verze Cassandry zase používaly surové nízkoúrovňové rozhraní API s názvem Thrift, novější verze nabízejí CQL, což je dotazovací jazyk podobný SQL, který se musí analyzovat a chápat serverem. Každý z nich se svým vlastním způsobem liší.

Libovolný nástroj nemá jen své vlastní výstřednosti, ale zároveň využívá zcela jinou filozofii a způsob vyjádření.

Neexistují žádné jednoduché metody, jak přejít mezi datovými sklady, a často musíte napsat doslova tuny propojovacího kódu, abyste získali jen možnost přejít v budoucnu.

Nemusí to být příliš těžké, když v systému používáte pár klíčů a hodnot, ale náročnost se složitostí může významně narůstat.

Nepříjemný fakt NoSQL Č. 6

Chybějící doplňky

Řekněme, že si nepřejete mít všechna data ve všech řádcích a chcete součet jednoho sloupce. Uživatelé SQL mohou položit dotaz prostřednictvím operace SUM a získat zpět jedno (jen jedno) číslo.

Uživatelé NoSQL ale musí dostat všechny údaje a součet si musí udělat sami. Samotný součet samozřejmě není problém, protože sečíst čísla trvá stejnou dobu na jakémkoliv počítači.

Přesun dat ale představuje poměrně pomalou záležitost a šířka pásma požadovaná k doručení všech těchto dat může být drahá.

NoSQL databáze mají několik doplňků. Chcete-li něco vykonat kromě ukládání či přečtení dat, budete to muset udělat sami. V mnoha případech to budete činit v jiném počítači s úplnou kopií dat.

Často ale může být užitečné, aby se všechny výpočty uskutečňovaly v počítači, na němž jsou zároveň uložená i data, protože jejich přenos by nějakou dobu trval. To ale může vadit.

Řešení NoSQL se však objevují stále častěji. Struktura dotazů Map a Reduce z databáze MongoDB umožňuje libovolnou konstrukci JavaScriptu pro zredukování dat.

Hadoop je zase výkonným mechanismem pro distribuci výpočtu do celé množiny strojů, které také uchovávají data.

Je to rychle se rozvíjející struktura, která nabízí soustavně se zlepšující nástroje pro vytváření sofistikované analýzy.

Je to velmi dobré, ale stále relativně nové. Technicky je Hadoop zcela jiný módní termín než NoSQL, přestože rozdíl mezi nimi mizí.

Nepříjemný fakt NoSQL Č. 7

Méně nástrojů

Samozřejmě můžete NoSQL umístit a spravovat na svém serveru. Můžete si také napsat vlastní kód, který by pracoval s databází – ukládal a načítal data.

Ale co když chcete dělat víc? Co když si chcete koupit jeden z oblíbených reportovacích balíků? Nebo řešení pro zobrazování grafů? Či si stáhnout nějaké open source nástroje pro vytváření grafů?

Omlouváme se, ale většina těchto řešení je vytvořená pro SQL databáze. Pokud chcete generovat reporty, tvořit grafy nebo dělat něco se všemi daty ve vaší NoSQL databázi, budete muset začít psát vlastní kód.

Standardní nástroje dokážou získávat data z řešení Oracle, Microsoft SQL, MySQL a Postgres.

Vaše data jsou v NoSQL? Na tom se teprve pracuje. A ještě nějakou dobu se na tom s velkou pravděpodobností pracovat bude.

I když se výrobci snaží vyhovět a dodat řešení pro jednu z databází NoSQL, budou muset při tvorbě pro další systém začínat zcela od počátku. Existuje více než 20 různých NoSQL variant, z nichž všechny mají vlastní filozofii a svůj způsob práce s daty.

Pro tvůrce zmíněných doplňkových nástrojů bylo dost těžké podporovat výstřednosti a nesrovnalosti v SQL, ale v oblasti NoSQL je tvorba takových nástrojů ještě výrazně složitější.

Tento problém ale bude pomalu mizet. Vývojáři mohou cítit nadšení z NoSQL a budou své nástroje upravovat, aby s těmito systémy dokázaly pracovat, ale bude to nějakou dobu trvat.

Mohou například začít s řešením MongoDB, ale vám to nijak nepomůže, pokud používáte Cassandru. V takových situacích pomáhají standardy, ale NoSQL na tom z hlediska standardizace také není nijak dobře…

Nevýhody NoSQL v kostce

Všechny tyto nevýhody NoSQL lze zredukovat na jednoduché prohlášení: Za rychlost NoSQL se zaplatilo obětováním funkcionality. Jestliže funkce nepotřebujete, budete spokojení, ale pokud je budete potřebovat v budoucnu, budete litovat.

Revoluce jsou pro kulturu technologií charakteristické. Nová skupina přijde a diví se, proč minulá generace vytvořila něco tak složitého, takže se vydá na cestu změn starých zvyklostí. Po chvíli si však začne uvědomovat, proč byly staré zvyklosti tak složité, a začne tyto funkce opětovně zavádět.

Vidíme to i ve světě NoSQL, jak některé z projektů začínají přidávat zpět vlastnosti, které vypadají jako transakce, schémata a standardy. To je ona povaha pokroku.

Bořímě věci, abychom je postavili znovu. Technologie NoSQL dokončila první fázi revoluce a nyní je čas na etapu druhou.

Král je mrtev. Ať žije král. ■

Big data usnadňují bezpečnostní korelace

Využití big dat pro potřeby zabezpečení dat může podnikům zajistit lepší povědomí o situaci, ale implementace může být poměrně obtížná.

JAIKUMAR VIJAYAN

Například RSA nebo Symantec nedávno představily své strategie nových přístupů pro agregaci, korelaci a analýzu dat, které podnikům pomohou prosít obrovské množství strukturovaných i nestrukturovaných dat a získat z nich relevantní výstupy týkající se různých bezpečnostních hrozeb.

Základní myšlenkou je, že takováto agregace a korelace dat firmám usnadní identifikovat trendy a hrozby, které by konvenční bezpečnostní nástroje nebyly obvykle schopné odhalit.

Na rozdíl od běžných ochranných systémů, které se zaměřují převážně na blokování útoků, nové přístupy dávají velmi vysoký důraz na detekci narušení, reakci a také na prevenci. Cílem je blokovat hrozby a zároveň umožnit detekci a reakci na takové, které už trvají nebo prošly ochrannou vrstvou.

Art Coviello, šéf RSA, si myslí, že nutnost použít pro tyto účely projekty spojené s big daty je výsledkem stále rostoucího počtu cílených a trvalých útoků APT (Advanced Persistent Attacks) na firmy a vládní organizace. Současný celkový objem a rozmanitost shromažďovaných a dolovaných dat spojených s bezpečností nutí přijmout nové přístupy k ochraně dat před nepřáteli.

„Namísto nasazování produktů pro jednotlivé body či hranice by podniky měly implementovat bezpečnostní model, který vychází ze skutečných hrozeb a má inteligenci, jež je dokáže rozpoznávat,“ vysvětluje Coviello.

Zatím nerovný souboj

Organizace se vtahují do stále více nerovné války s kyber zločinci, kteří jsou lépe vyzbrojeni, připraveni a organizováni než ony samotné, upozorňuje Francis deSouza, prezident produktů a služeb v Symanteku.

„Útočníkům stačí uspět jen jednou, ale my musíme uspět vždy,“ připomíná jednoduchou rovnici deSouza. Podle něj by se spíše než čistě na blokování všech hrozeb měly organizace zaměřit na použití analytických řešení typu big dat, která dokážou různé metody vniknutí zjistit, a tak jejich dopad zmírnit.

Z teoretického hlediska je myšlenka posílení bezpečnosti prostřednictvím zkoumání a analýz velkých datových objemů dobrá, domnívá se řada IT manažerů i bezpečnostních expertů.

Dosáhnout tohoto stavu však může být dosti namáhavé, varuje Christopher Pierson, šéf ochrany dat a dodržování předpisů ve společnosti LSQ Holdings, která poskytuje finanční služby. *„Myslím, že problém spočívající v nutnosti mít vhled do různých logů a všech bezpečnostních zařízení je dlouhodobý,“* prohlašuje Pierson.

Současné nástroje SIEM (Security Incident and Event Management, správa událostí a incidentů zabezpečení) již umožňují agregovat obrovské množství protokolů z různých bezpečnostních zařízení a přenášet je všechny do jednoho analytického systému.

Skutečný problém SIEM je ale schopnost tyto údaje analyzovat a korelovat tak, aby se včas zjistily pokusy o prolomení, detekovaly skutečné útoky a udělaly následné vhodné protiakce.

Korelace klíčem k úspěchu

Agregace je přitom jen jednou částí problematiky. Zcela jinou záležitostí však je schopnost najít v takto získaných údajích smysl. Klíčem k situačnímu povědomí jsou korelační pravidla a procesy, které organizace zavedla pro analýzu dat a podle nichž se dá efektivním způsobem reagovat.

„Velkou výzvou řešení big dat je získání extrémně užitečných informací,“ prohlašuje Andrew Wild, šéf bezpečnosti ve firmě Qualys, která se na řešení SIEM specializuje. Problém, kterému mnohé podniky čelí, podle něj nespočívá ani tak v nedostatku dat, jako spíše v absenci vhodného způsobu jejich využití.

„Směrovače, přepínače a další komunikační prvky – všechny mohou sledovat okolní dění. Tyto produkty vědí, jaké pakety procházejí sítí. Problém ale je v tom, že všechny tuto údaje existují v různých úložištích, která se obvykle vůbec neintegrují,“ prohlašuje Wild.

Zmíněné nástroje tak nejsou jednotlivě schopné poskytnout samy o sobě mnoho užitečných informací, takže velkým úkolem pro big data je najít způsob, jak údaje shromažďovat a získat z nich vysokou přidanou hodnotu.

„Big data jsou, pokud jde o bezpečnost, velkou výzvou,“ tvrdí Jerry Sto. Tomas, šéf globálního zabezpečení informací ve společnosti Allergan.

Podle něj spousta logů, které podniky shromažďují, existuje odděleně a navíc z bezpečnostního hlediska často jde o soubor obtížně využitelných údajů.

Napraví tuto situaci projekty využívající big data? Zřejmě ano – a nemusí to tak dlouho trvat. ■

SEZNAM INZERUJÍCÍCH FIREM

DNS 8, 9
www.dns.cz

EMC Czech Republic 7
www.emc.com

HEWLETT-PACKARD ..... 4. obálka
www.hp.cz

SAP ČR 11
www.sap.com/cz

SAS Institute ČR 12, 13
www.sas.cz

Teradata Česká republika 22, 23
www.teradata.com

T-Systems Czech Republic 20, 21
www.t-systems.cz

Větší data.
Lepší výsledky.

HP Vertica je účelově vyvinutá platforma pro analýzu velkých objemů dat (Big Data) v reálném čase.

Analýzy v reálném čase
50× až 1 000× rychlejší oproti tradičním databázím

Otevřená architektura
Podpora Hadoop, programovacího jazyka R a široké nabídky dalších BI a ETL nástrojů

Optimalizované ukládání
10× až 30× více dat na jednom serveru

Masivní škálovatelnost
Prakticky neomezený počet uzlů pro zpracování až petabajtů dat

Flexibilita
Možnost nasazení na standardizovaný hardware nebo do prostředí privátního/veřejného cloudu

Podívejte se, proč platformě Vertica důvěřují tisíce zákazníků.
Zaregistrujte se a získejte 30ti denní trial verzi zdarma na webu vertica.com/evaluate